Les usages de l'intelligence artificielle

Edition 2018

Novembre 2018

Olivier Ezratty

A propos de l'auteur

Olivier Ezratty — olivier@oezratty.net , http://www.oezratty.net , @olivez

consultant et auteur — +33 6 67 37 92 41

Olivier Ezratty conseille les entreprises dans l'élaboration de leurs stratégies produits et marketing, avec une focalisation sur les projets à fort contenu technique et scientifique (objets connectés, intelligence artificielle, medtechs, biotechs, …). Il leur apporte un triple regard : technologique, marketing et management ainsi que la connaissance des écosystèmes dans les industries numériques.

Il a réalisé depuis 2005 des missions diverses d'accompagnement stratégique, de conférences ou formations dans différents secteurs d'activité tels que les **médias/télécoms** (Orange, Bouygues Télécom, TDF, Médiamétrie, BVA), la **finance et l'assurance** (groupe BPCE, Caisse des Dépôts, Crédit Agricole, Crédit Mutuel-CIC, Generali, MAIF, Société Générale) ainsi que l'**industrie et les services** (Schneider, Camfil, Vinci, NTN-STR, Conseil Supérieur du Notariat, Ordres des Experts Comptables).

Olivier Ezratty est l'auteur du **Rapport du CES de Las Vegas**, publié à la fin janvier de chaque année depuis 2006, du **Guide des Startups** qui permet aux entrepreneurs de disposer d'une mine d'information pour créer leur startup avec plus de 300 000 téléchargements à date, mis à jour tous les ans (22e édition et 12e année en 2018), ainsi que des ebooks **Les usages de l'intelligence artificielle** en octobre 2017 et novembre 2018 et **Comprendre l'informatique quantique** en septembre 2018, destinés à vulgariser ces deux vagues technologiques auprès des entreprises. Le tout est publié sur le blog « Opinions Libres » (http://www.oezratty.net) qui traite de l'innovation technologique vue sous les angles scientifiques, technologiques, entrepreneuriaux et des politiques publiques de l'innovation.

Ses contributions s'appuient sur un rôle actif dans l'écosystème de l'innovation et sous différentes casquettes, notamment dans l'univers des startups :

- Expert auprès de l'accélérateur **Wilco** – ex Scientipôle Initiative – depuis 2007.
- Formateur sur l'intelligence artificielle auprès de **Cap Gemini Institut** depuis 2017.
- Membre depuis fin 2015 du Comité Stratégique, puis Scientifique de l'**ARCEP**.
- Membre du jury de divers **concours entrepreneuriaux** comme le Grand Prix de l'Innovation de la Ville de Paris ou la Startup Academy, mentor dans de nombreux **Startups Weekends**.

Il intervient comme conférencier dans divers établissements d'enseignement supérieur tels que HEC, Neoma Rouen, SciencePo, CentraleSupelec, Télécom Paristech ou Les Gobelins, sur le marketing de l'innovation dans les industries numériques, sur l'entrepreunariat et le product management, en français comme en anglais selon les besoins. Olivier Ezratty est aussi le co-initiateur en 2012 de « Quelques Femmes du Numérique ! » (http://www.qfdn.net), devenu une association en 2016, et qui vise à augmenter la place des femmes dans les métiers du numérique, en sensibilisant les jeunes à ces métiers. Il y contribue comme photographe amateur et a réalisé à la mi 2018 plus de 700 portraits de femmes exercant des métiers dans le numérique, notamment à des rôles techniques et entrepreneuriaux.

Et avant tout cela ? Olivier Ezratty débute en 1985 chez **Sogitec**, une filiale du groupe Dassault, où il est successivement Ingénieur Logiciel, puis Responsable du Service Etudes dans la Division Communication. Il initialise des développements sous Windows 1.0 dans le domaine de l'informatique éditoriale ainsi que sur SGML, l'ancêtre de HTML et XML. Entrant chez **Microsoft France** en 1990, il y acquiert une expérience dans de nombreux domaines du mix marketing : produits, canaux, marchés et communication. Il lance la première version de Visual Basic en 1991 ainsi que Windows NT en 1993. En 1998, il devient Directeur Marketing et Communication de Microsoft France et en 2001, de la Division Développeurs dont il assure la création en France pour y lancer notamment la plateforme .NET et promouvoir la plate-forme de l'éditeur auprès des développeurs, dans l'enseignement supérieur et la recherche ainsi qu'auprès des startups. Olivier Ezratty est ingénieur Centrale Paris (1985).

Ce document vous est fourni à titre gracieux et est sous licence « Creative Commons » dans la variante « Paternité-Pas d'Utilisation Commerciale-Pas de Modification 2.0 France ».

Voir http://creativecommons.org/licenses/by-nc-nd/2.0/fr/

Photo de couverture : schéma trouvé ici et modifié.

Table des matières

Objectifs et contenu

Depuis 2015, l'intelligence artificielle (IA) est devenue l'une des grandes priorités des grands acteurs de l'économie numérique et des entreprises. La vague de l'IA commence à égaler celles de l'Internet et de la mobilité en termes d'impact. Il est plutôt difficile d'y échapper ! C'est devenu un outil de la compétitivité des entreprises aussi bien des secteurs du numérique que du reste de l'économie.

L'IA est aussi devenue rapidement un sujet de société et un objet politique. Elle génère son lot de questions sur le futur du travail, sur l'éthique, sur la répartition des richesses, sur la souveraineté économique et même sur le devenir de l'espèce humaine comme cela a été souligné entre autres dans le Rapport de la Mission Villani publié en mars 2018.

Cette vague technologique est atypique dans plusieurs dimensions. Les technologies et méthodes de l'IA sont méconnues ou assimilées à tors et à travers, y compris par la majorité des professionnels du numérique, d'où la propagation de nombreux mythes à son sujet, un peu trop directement inspirés par la science-fiction ou par les thèses singularistes[1]. On croit par exemple que le « deep learning » ou « apprentissage profond » raisonne alors qu'il ne fait qu'appliquer, en général, des modèles probabilistes. De même, dans le machine learning, l'apprentissage non-supervisé n'est pas de l'apprentissage supervisé autonome. La sémantique de l'IA est donc porteuse de contre-sens qu'il faut décrire et éviter. Le sens même de l'appellation « IA » est sujet à d'infinis débats, trop connotés d'accents anthropomorphiques alors que l'IA est artificielle et donc différente de l'intelligence humaine même si les deux ont quelques zones de recouvrement.

Nombre des évangélistes de l'IA la vulgarisent au point de saupoudrer leurs propos d'amalgames, d'approximations, de revues de presse pas toujours bien vérifiées et d'exagérations exaspérantes pour ceux qui cherchent à comprendre les technologies de près et à prendre une certaine distance vis-à-vis des effets d'annonces[2].

Les prophètes de mauvais augure voient de leur côté arriver à grand pas une intelligence artificielle généralisée qui rendrait l'Homme caduque, la connexion directe des IA aux cerveaux pour mieux la contrôler ou se faire hacker par elle et autres délires singularistes et transhumanistes invérifiables tellement ils sont éloignés dans le temps si ce n'est quasiment impossibles à réaliser. Tout ceci vient d'une lecture bien trop anthropomorphique des techniques de l'IA alors que l'on ne sait même pas décrire avec précision le fonctionnement de l'intelligence et de la conscience humaines ! Et dire que tout cela ne faisait pas du tout partie des intentions du créateur de l'expression d'IA, John MacCarthy, en 1955 !

Cela a pour conséquence de saturer l'espace médiatique depuis 2015 avec des propos qui se focalisent sur le pour et le contre de l'IA plutôt que sur ses aspects tangibles allant des techniques de l'IA à ses applications. Or l'IA est devenue un sujet central pour les entreprises, dans la lignée des autres grandes vagues technologiques du numérique, comme celles de la mobilité et du cloud et avant celles, plus incertaines, de la Blockchain ou de l'informatique quantique.

Cet ouvrage se positionne à contre-courant de cette littérature anxyogène. Il vise à démythifier de manière posée l'état de l'art de l'IA et à rendre son adoption et son utilisation aussi pratiques que possible pour les entreprises.

---

[1] Mythes que j'ai eu l'occasion de décrire en septembre 2017 dans Douze mythes de l'intelligence artificielle.

[2] Ils utilisent des techniques de prise de parole qui consistent à forcer le trait au point de travestir la réalité, comme prétendre qu'une solution qui relève de la prospective ou du concept est déjà opérationnelle, sans d'ailleurs forcément s'en rendre compte eux-mêmes ! Le fact-checking demande des efforts particuliers dans le domaine de l'intelligence artificielle.

Ce document se veut être une sorte de grande boîte à idées pour les entreprises qui se demandent où et comment intégrer l'intelligence artificielle dans leur système d'information.

Cet ebook est en fait le troisième d'une série lancée en 2016.

La **première édition**, Les avancées de l'intelligence artificielle, était une compilation de neuf articles publiés entre mars et mai 2016 sur mon blog Opinions Libres (197 pages). Elle débroussaillait à un premier niveau le brouhaha de l'époque sur l'IA. J'y décrivais la manière dont l'écosystème de l'IA se construisait à l'époque. Le contenu technique y était encore approximatif.

La **seconde édition** Les usages de l'intelligence artificielle d'octobre 2017 (362 pages) complétait la précédente avec un contenu technique plus dense et précis, corrigeant au passage un bon nombre d'erreurs ou d'appréciations qui émaillaient la première édition. On y trouvait une partie dense sur le matériel de l'IA. Cette édition décrivait aussi les usages de l'IA par secteur d'activité, avec une douzaine de marchés comme les transports, la santé, la finance, l'assurance, la distribution et les médias. Enfin, le document décrivait la structure de l'offre en IA avec ses outils de développement, ses startups et ses grandes entreprises numériques, souvent américaines. Cette édition m'a aussi permis de structurer mes différentes conférences et formations sur l'intelligence artificielle destinées aux entreprises, notamment le cours de deux jours que je délivre chez Cap Gemini Institut depuis novembre 2017.

Cette **troisième édition** 2018 est une très importante mise à jour de la seconde. En un an, nombre de choses ont évolué. Les entreprises ont commencé à prendre en main l'IA. Les études de cas sont plus nombreuses et commencent à être mieux documentées. Les stratégies des acteurs ont évolué. Les pouvoirs publics se sont emparés du sujet avec le pinacle du rapport de la Mission Villani en mars 2018. On n'a jamais autant évoqué l'IA dans les médias et même dans la fiction ! Les modifications de cette version sont omniprésentes dans le texte. Quelques nouveaux marchés verticaux ont été ajoutés. J'ai aussi réintégré une part du contenu de divers articles que j'ai pu publier entre octobre 2017 et avril 2018, sur la dimension créative de l'IA, sur l'IA émotionnelle, sur l'IA et la cybersécurité et sur l'IA symbolique.

L'ambition de ce document est de revenir au présent et au futur proche pour comprendre les usages et techniques de l'IA dans les entreprises et les aider à en tirer le meilleur parti. Il s'agit de mettre les fantasmes de côté et de rentrer dans le concret.

L'intelligence artificielle n'est pas un produit. Elle ne se présente généralement pas sous la forme de logiciels packagés traditionnellement comme un traitement de texte, une application mobile ou un système d'exploitation. Il n'y a pas de logiciels d'intelligence artificielle mais *des* solutions logicielles et matérielles qui exploitent des briques d'intelligence artificielle variées qui s'appuient sur plusieurs dizaines de briques logicielles différentes qui vont de la captation des sens, notamment audio et visuel, à l'interprétation des informations, au traitement du langage et à l'exploitation de grandes bases de données et de connaissances structurées ou non structurées. Leur création et intégration est encore une affaire de bricolage et de tâtonnements, si ce n'est d'un véritable artisanat.

Nous en sommes toujours à l'âge de pierre, avec seulement une soixantaine d'années de recul sur la question et un peu plus d'une vingtaine d'années pour ce qui est du deep learning. Mais les chercheurs du domaine le font rapidement avancer et le passage de la recherche à la production est de plus en plus rapide, les outils de développement de l'IA permettant de les mettre en pratique assez facilement pour peu que les bons jeux de données soient disponibles. Or les jeux de données sont nombreux dans les entreprises ou en open data pour entraîner ses modèles !

Dans mes différentes lectures scientifiques, en particulier autour du calcul quantique, j'ai découvert aussi la réalité des limites scientifiques de l'informatique. Celle-ci ne peut pas résoudre tous les problèmes qui se présentent. Il existe des problèmes trop complexes pour les ordinateurs, même quantiques. Des problèmes sont dits « indécidables » car ils n'ont pas de solution, ou tout du moins pas de solutions parfaites. L'IA peut aider à trouver des solutions optimales ou non et c'est au libre arbitre de l'Homme de les choisir.

Cela permet de mieux comprendre pourquoi l'IA surhumaine qui pourrait tout prévoir, calculer et créer est un mythe. Mais l'IA d'aujourd'hui peut servir d'accélérateur de l'intelligence humaine.

L'IA est un grand tonneau des Danaïdes scientifique. On n'arrive jamais à tout comprendre et à tout appréhender des techniques et domaines d'applications de l'IA. Chercher un "expert en IA"[3] revient maintenant à demander "un expert en logiciels" ou un "expert en informatique" sans compter le top avec "l'expert en transformation digitale".

Contrairement à un lieu commun répandu, les techniques et méthodes de l'IA évoluent sans cesse. Ce n'est pas qu'une question de puissance de machine ou de volume de données. Certains vieux de la vieille de l'IA considèrent qu'elle évolue très lentement et que rien n'a changé depuis leurs études. Ca me semble bien exagéré, même après avoir dé-marketinguisé les progrès récents de l'IA.

Selon certains, il faudrait un doctorat en IA pour pouvoir créer une solution intégrant de l'IA. C'est peut-être vrai aujourd'hui dans certains cas mais de nombreux outils de développement et d'intégration arrivent sur le marché qui permettent à des développeurs moins qualifiés, voire même à des cadres, de créer eux-mêmes des solutions intégrant des briques d'IA. C'est avec sa démocratisation que l'on peut évaluer l'évolution de la maturité d'une technologie donnée.

Je ne vais pas vous faire croire que j'ai tout compris à l'IA. Ce n'est pas le cas et des domaines de l'IA m'échappent encore. J'ai plein d'interrogations diverses, en particulier autour des techniques de traitement du langage et de leur dimensionnement[4] ! J'explore aussi les évolutions du champ de la représentation des connaissances et du raisonnement automatique, l'un des domaines les plus ardus de l'IA[5].

Comme dans sa seconde version, cet ebook adopte un découpage en parties qui est adapté à la compréhension à la fois des techniques de l'IA, de ses outils de développement et, surtout, de ses usages dans les entreprises.

Ce document est organisé en huit grandes parties encore mieux segmentées que dans la version précédente :

- **Histoire et sémantique de l'IA :** qu'est-ce que l'IA ? Qui a créé la discipline ? D'où vient cette appellation ? Pourquoi personne n'est d'accord sur le sens qu'il faut lui donner ? Comment l'IA est-elle segmentée d'un point de vue technique ? Quels sont ses grands courants intellectuels ? Comment cette discipline nouvelle a-t-elle progressé depuis les années 1950 ? Pourquoi a-t-elle connu deux grands hivers et qu'est-ce qui explique la dynamique actuelle ? Est-elle durable ? Où en est-on aujourd'hui ? Comment l'IA se compare-t-elle à l'intelligence humaine ?
- **Algorithmes et logiciels de l'IA** : quelles sont les principales briques mathématiques et algorithmiques de l'IA ? Le raisonnement automatique et les systèmes experts et pourquoi en parle-t-on moins que pendant les années 1980 ? Quelles sont les techniques et applications du machine learning, des réseaux de neurones et du deep learning ? Les progrès récents viennent-ils du logiciel, du matériel ou des données ? Quels sont les outils de développement et de création d'applications de l'IA et pourquoi la majorité sont-ils open source ? Comment les briques d'intelligence artificielle progressent-elles ? Quid de l'intelligence artificielle généralisée ? Est-ce un fantasme ? Peut-on facilement reproduire le fonctionnement du cerveau humain ? Quels sont les projets allant dans ce sens et peuvent-ils aboutir ?
- **Données de l'IA** : quel est le rôle des données dans l'IA ? D'où viennent-elles ? Quelles sont les données ouvertes exploitables par l'IA ? Qu'est-ce que le biais des données dans l'IA et comment l'évite-t-on ? Quels sont les capteurs qui alimentent les données de l'IA ?

---

[3] Voir Confession of a so-called AI expert de Chip Huyen, juillet 2017.

[4] Comme comprendre et expliquer dans le détail le fonctionnement des réseaux de neurones à mémoire de type LSTM.

[5] Voir Que devient l'IA symbolique ?, que j'ai publié en avril 2018.

- **Matériel de l'IA** : quelles sont les ressources matérielles qui font avancer l'IA ? Comment évolue la loi de Moore ? Quel est le rôle des GPU et des processeurs neuromorphiques dans l'IA ? Comment se distinguent-ils et comment les classifier ? Quels sont les acteurs de ce marché ? Pourquoi il y-a-t-il une grande différence entre l'entraînement d'une IA et son exécution dans la consommation de ressources matérielles ? Est-ce que l'informatique optique et quantique auront un impact sur l'IA ? Comment sont gérées les ressources en cloud de l'IA ainsi que du côté des systèmes embarqués ? Comment architecturer les solutions d'IA en tenant compte des évolutions des processeurs, des télécommunications, des questions énergétiques et de sécurité ?
- **Applications génériques de l'IA** : quelles sont les applications génériques et horizontales de l'IA, dans le traitement de l'image, du langage, dans la robotique, dans le marketing, les ressources humaines, la comptabilité ainsi que dans la cybersécurité ?
- **Applications métiers de l'IA** : quelles sont les grandes applications et études de cas de l'IA selon les marchés verticaux comme les transports, la santé, la finance, l'assurance, l'industrie, la distribution, les médias, le tourisme, l'agriculture, les métiers juridiques, les services publics, la défense et le renseignement ? S'y ajoutent dans cette édition : les utilities, l'éducation, le BTP et l'immobilier, le luxe, les services et le conseil et l'Internet générique. Pourquoi certains de ces marchés sont plus dynamiques que d'autres ? Comment les startups permettent aux entreprises d'innover dans ces différents marchés[6] ?
- **Acteurs de l'IA** : quelle est la stratégie et quelles sont les offres en IA des GAFAMI étendus, dont IBM, Google, Microsoft, Facebook, SalesForce, Oracle et plein d'autres encore ? Quid des Chinois ? Comment certains de ces acteurs se déploient-ils de manière verticale ? Comment se développent les startups en général et puis celles de l'écosystème français en particulier ? Comment évaluer la valeur ajoutée en IA des startups et autres acteurs de l'écosystème ? Comment les solutions d'IA sont-elles commercialisées ? Quelle est la part qui relève de produits et celle qui dépend des services et des données ?
- **IA et société** : les points de vue et études sur l'impact potentiel de l'IA sur l'emploi, les métiers et sur la société en général. Quelles sont les limites des prédictions ? Comment éviter de se faire robotiser ? Comment se préparer au niveau des compétences ? Quelles sont les grandes lignes de l'impact de l'IA sur la politique et les politiques de l'IA en France et ailleurs dans le monde ? Quel est l'état de la géopolitique de l'IA ? La Chine va-t-elle nous envahir avec son IA ?
- **IA et entreprise** : comment les entreprises peuvent-elles intégrer l'IA dans leur stratégie ? Quelles sont les bonnes méthodes et pratiques ? Comment gérer les compétences ? Comment benchmarker les solutions d'IA ? Comment s'organiser ? Comment intégrer l'IA dans les autres dynamiques d'innovations liées au numérique ? Comment va évoluer le métier de développeur ? Comment se former en général ?

Ce document s'appuie à la fois sur des rencontres avec des chercheurs, entrepreneurs et entreprises et surtout, sur une recherche bibliographique extensive. La littérature disponible sur le sujet est abondante, notamment les excellents cours de nombreuses universités comme ceux de Stanford ou Berkeley mais aussi ceux du Collège de France, avec et au-delà de Yann LeCun. C'est la magie d'Internet quand on prend le temps de creuser ! L'abondance de documentation scientifique permet notamment de faire le pont entre les effets d'annonces et la réalité sous-jacente.

Je remercie au passage les relecteurs de cet ebook qui ont permis d'en parfaire le contenu : **Benoît Bergeret**, **Françoise Soulié Fogelman**, **Antoine Couret** et **Dimitri Carbonnelle**.

Bonne lecture !

Olivier Ezratty, 15 novembre 2018

[6] Je cite un très grand nombre de startups dans ce document. Il se peut que telle ou telle startup soit en déclin, ait fait un pivot ou n'existe plus. C'est la vie habituelle des startups. Je corrige le document au fil de l'eau lorsque nécessaire. Prévenez moi !

Histoire et sémantique de l'IA

L'intelligence artificielle génère toutes sortes de fantasmes pour les uns et de craintes pour les autres. Elle s'inscrit dans une Histoire humaine faite de machines qui démultiplient la puissance humaine, d'abord mécanique, puis intellectuelle avec de l'automatisation et une recherche effrenée de puissance et d'effets de levier. Elle s'alimente de nombreux mythes, de celui du golem qui mène aux robots ou Frankenstein et à celui d'Icare[7].

Tout cela alors qu'il n'existe même pas de consensus sur la définition de ce qu'est l'intelligence artificielle ! Elle est source d'une interminable bataille sémantique[8] qui touche notamment les start-tups du numérique[9]. Cela s'explique notamment par la difficulté que l'Homme a pour définir sa propre intelligence, sa conscience et ses mécanismes de raisonnement. Nous ne savons pas encore tout du fonctionnement du cerveau biologique. Mais il est vain de vouloir le copier. Il doit servir de source d'inspiration pour faire différemment et mieux, tout comme l'avion a des caractéristiques que n'ont pas les oiseaux.

Cohabitent donc des définitions étroites et larges de l'IA. Pour certains, seul le deep learning est digne de faire partie de l'IA et le machine learning et même les moteurs de règles, pas du tout. Comme si seules les technologies un peu magiques dans leur apparence pouvaient faire partie de l'IA.

Pour nombre de spécialistes du secteur et votre serviteur, toutes les technologies de l'histoire de l'IA en font partie[10]. Certaines ont une dimension anthropomorphique comme la vision artificielle ou les agents conversationnels, d'autres, beaucoup moins lorsqu'elles analysent de gros volumes de données pour identifier des corrélations, des tendances ou faire des prévisions. Des composantes isolées de l'intelligence humaine sont déjà intégrées dans les machines avec des capacités de calcul brutes et de mémoire qui dépassent celles de l'Homme depuis des décennies. On associe aussi à l'IA les logiciels qui exploitent la force brute pour gagner à divers jeux de société.

On peut opérer un jugement de Salomon en rappelant qu'artificiel est le contraire de naturel. Le débat sur la "vraie" IA est sans fin. Il n'y aura probablement jamais d'IA exactement équivalente à l'intelligence humaine, ne serait-ce que parce que celle-ci est construite autour d'un substrat biologique qui l'alimente avec son cerveau, ses sens, son système hormonal, ses muscles, son squelette, ses hormones, ses besoins vitaux, une relation au temps et un processus de développement bien particulier. Le cerveau est lent mais massivement connecté et parallèle.

Plus des deux tiers des neurones de notre cerveau sont dans le cervelet alors qu'il n'en représente que 10% du poids. Celui-ci cogère les mouvements appris avec le cortex. Aussi curieux que cela puisse paraître, d'un point de vue quantitatif, une bonne part de notre intelligence est située dans la capacité à se mouvoir dans l'espace, et que les robots ont d'ailleurs bien du mal à imiter.

---

[7] Voir Tout ce que vous pensez savoir sur l'intelligence artificielle est complètement faux, où tout n'est pas forcément vrai, de Mathilde Rochefort, septembre 2016.

[8] Voir notamment « Intelligence artificielle – vers une domination programmée ? », de Jean-Gabriel Ganascia, seconde édition publiée 2017 d'une première édition datant de 1993, qui raconte très bien les débuts et le parcours de l'IA comme science.

[9] La querelle sémantique qui atteint l'univers des startups est toujours d'actualité. Celles-ci feraient de « l'IA washing », peignant aux couleurs de l'IA des solutions qui n'en contiennent pas forcément. Réflexion faite, cette notion d'IA washing est exagérée. Ce n'est pas parce que certaines utilisent des briques technologiques prêtes à l'emploi qu'elles ne font pas d'IA ou que leur solution n'intègre pas d'IA. C'est un peu comme si on disait qu'un site web réalisé en Wordpress avec un thème standard au lieu d'être développé avec son propre framework en Ruby on Rails avec un front-end personnalisé en React ou Angular n'était pas "de l'Internet". Reste à définir "une IA", qui est toujours un assemblage de plusieurs composantes (data, algos, hard, savoir faire métier) et à ausculter les startups en examinant le CV de leurs équipes techniques, leur solution et les données qui les alimentent. Ce qui permet de faire un premier tri.

[10] Voir cette intéressante histoire de l'IA : The Quest for Artificial Intelligence - A history of ideas and achievements de Nils Nilsson, Stanford University (707 pages). Un ouvrage qui date d'avant la grande vague du deep learning.

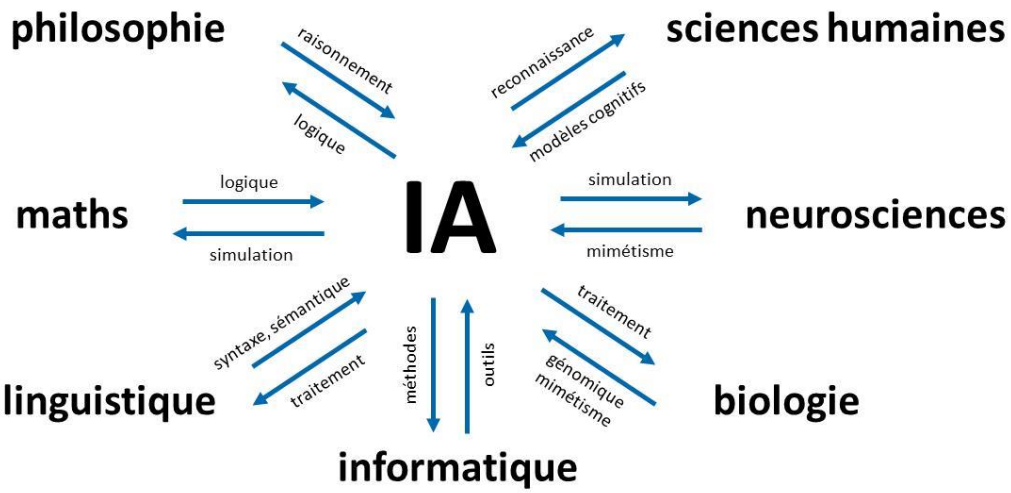
La partie cognitive du cerveau ne représente qu'à peine 10% des neurones du cerveau. C'est surprenant [11]! Cela n'enlève rien à l'intelligence humaine qui est plus que diverse.

D'un point de vue sémantique, il faut peut-être s'échapper de la définition anthropomorphique de l'intelligence. Il y a des intelligences diverses. La notion d'intelligence intègre la capacité à comprendre, apprendre et à s'adapter à des situations nouvelles. On retrouve à des niveaux variables ces capacités dans des systèmes non vivants ou des systèmes vivants très simplifiés par rapport au cerveau humain. Les fourmis n'ont que 250 000 neurones en tout mais font preuve d'une certaine intelligence collective.

L'intelligence de l'Homme est très diverse, entre le pékin moyen et le scientifique de génie et entre ma médiocre maîtrise de la flûte à bec et les meilleurs joueurs de violon du monde, Jimmy Page et Jimmy Hendrix à la guitare !

Rien n'empêche d'ailleurs de créer d'autres formes d'intelligences que l'intelligence humaine. Les machines démuliplient la force physique de l'Homme. A charge pour l'IA de démultiplier l'intelligence humaine. Cela devrait être une boîte à outils pour l'Homme.

L'intelligence artificielle est une discipline de l'informatique intimement liée à d'autres sciences : les mathématiques, la logique et les statistiques qui lui servent de base théorique, les sciences humaines (sciences cognitives, psychologie, philosophie, linguistique, …) et la neurobiologie qui aident à reproduire des composantes de l'intelligence humaine par biomimétisme, et enfin, les technologies matérielles qui servent de support physique à l'exécution des logiciels d'IA.



L'IA représente à la fois un pan entier de l'informatique avec sa diversité, ses briques technologiques, ses méthodes, ses assemblages et solutions en tout genre, et un ensemble de technologies qui sont maintenant imbriquées dans quasiment tous les pans du numérique. C'est un véritable écosystème hétéroclite. La grande majorité des solutions commerciales d'IA sont faites de bric et de broc, en fonction de besoins spécifiques.

On est loin d'avoir sous la main des solutions d'IA génériques, tout du moins dans les entreprises. Dans le grand public, des briques d'IA sont déjà utilisées au quotidien sans forcément que les utilisateurs le remarquent. C'est par exemple le cas des systèmes de suivi du visage dans la mise au point des photos et vidéos dans les smartphones.

Les briques technologiques les plus génériques de l'IA sont les bibliothèques logicielles et outils de développement comme **TensorFlow**, **PyTorch**, **scikit-learn** ou **Keras**[12].

[11] C'est le paradoxe de Moravec selon lequel les facultés humaines les plus difficiles à reproduire sont celles qui sont en apparence les plus simples pour l'Homme, à savoir ses capacités motrices.

Hauts et bas de l'IA

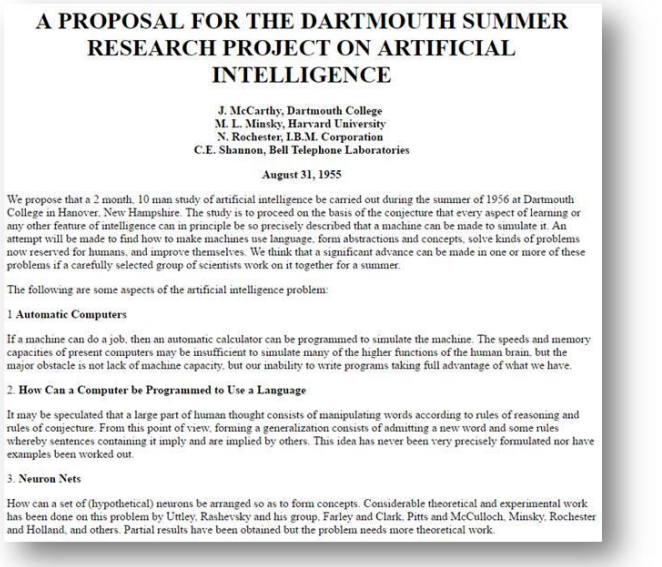
L'IA puise ses sources dans le concept théorique de *calculus ratiocinator* de **Gottfried Wilhelm Leibnitz** (circa 1671), la machine et le fameux test de **Turing** que l'on ne présente plus (1935 et 1950), les neurones formels de **Warren McCulloch** et **Walter Pitts** (1943), l'architecture de **John Von Neuman** qui sert encore de base à celle des ordinateurs traditionnels d'aujourd'hui (1945) ou encore le théorème de l'information de **Claude Shannon** (1949).

L'histoire moderne de l'intelligence artificielle a cependant véritablement démarrée au moment du **Summer Camp de Darmouth**, organisé entre le 18 juin et le 17 août 1956 à Hanover dans le New Hampshire aux USA.

Le groupe de travail fondateur du summer camp de Darmouth comprenait **John McCarthy**, **Claude Shannon**, à l'époque au MIT, **Marvin Minsky**, à l'époque à Princeton, **Allan Newell** et **Herbert Simon** de Carnegie Tech, et **Arthur Samuel** et **Nathaniel Rochester**, tous deux d'IBM. Ce dernier avait conçu l'IBM 701, le premier ordinateur scientifique d'IBM lancé en 1952.

congrès de Darmouth – USA – 1956

définit le périmètre d'investigation de l'IA



A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1. **Automatic Computers**

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

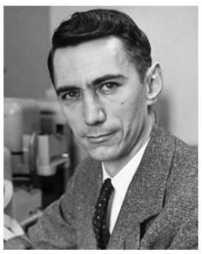
2. **How Can a Computer be Programmed to Use a Language**

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalization consists of admitting a new word and some rules whereby sentences containing it imply and are implied by others. This idea has never been very precisely formulated nor have examples been worked out.

3. **Neuron Nets**

How can a set of (hypothetical) neurons be arranged so as to form concepts. Considerable theoretical and experimental work has been done on this problem by Uttley, Rashevsky and his group, Farley and Clark, Pitts and McCulloch, Minsky, Rochester and Holland, and others. Partial results have been obtained but the problem needs more theoretical work.

Marvin Minksy
1927-2016



Claude Shannon
1916-2001

John McCarthy
1927-2011

Nathaniel Rochester
1919-2001

Il s'agissait d'une sorte de hackathon intellectuel de huit semaines dont l'objectif était de réfléchir aux concepts permettant de reproduire dans des machines diverses composantes de l'intelligence humaine de base comme la maîtrise du langage, la vision et le raisonnement[13]. Environ une trentaine de personnes y ont participé en tout.

Plaçons-nous dans le contexte de l'époque : l'informatique est alors un marché naissant avec à peine quelques centaines d'ordinateurs dans le monde, fonctionnant tous avec des lampes. 1955 était l'année de l'apparition des premiers mainframes à transistors avec notamment l'IBM 702 dont seulement 14 exemplaires ont été fabriqués.

---

[12] Voir à ce sujet Software 2.0 d'Andrej Karpathy, novembre 2017, qui banalise l'usage de l'IA via ses outils de développement.

[13] Précisément : "*The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it [...] Topics to study: automatic computers and programs for them; programming computers to use a language; neuron nets; machine self-improvement; classifying abstractions; and two areas particularly relevant to Solomonoff's work, though they would not be part of AI for many years — how to measure the complexity of calculations, and randomness and creativity*".

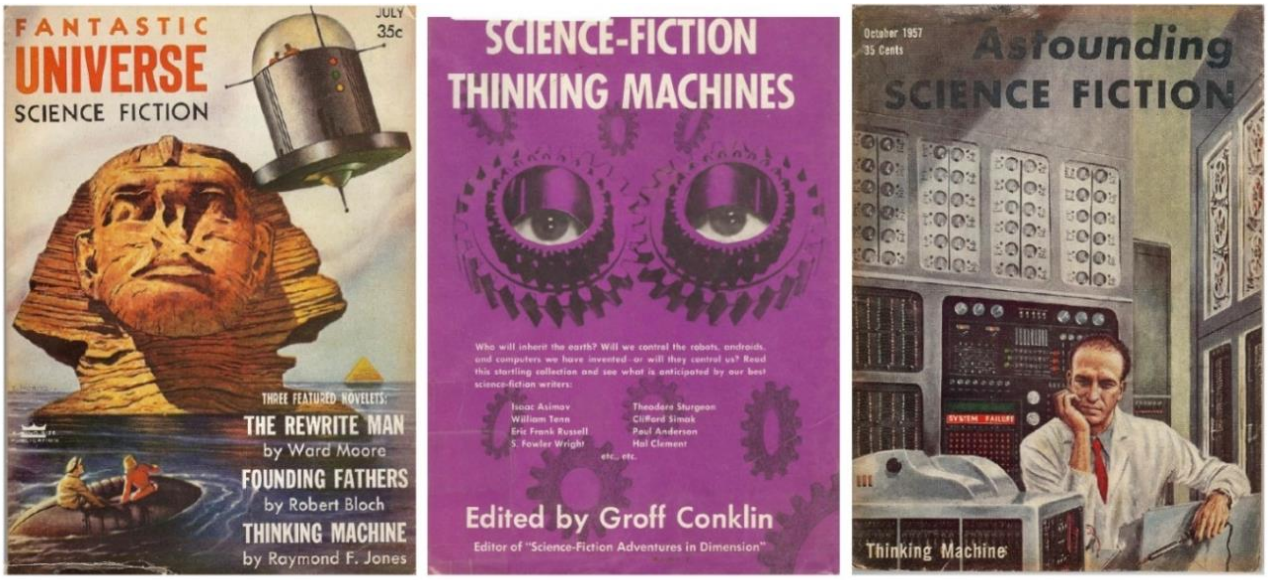
C'est l'IBM 360 qui va véritablement faire émerger le marché des mainframes. Il sera commercialisé seulement 8 ans plus tard, en 1964 !

Voyage éternel ou aboutissement ?

L'expression « intelligence artificielle » fut couchée sur papier le 31 août 1955 par l'un des initiateurs de ce summer camp, **John McCarthy**[14] dans la note de 13 pages proposant l'organisation du Summer Camp de Darmouth[15].

Elle recouvre les sciences et technologies qui permettent d'imiter, d'étendre et/ou d'augmenter l'intelligence humaine avec des machines. Une autre définition courante de l'IA est le champ académique de la création de logiciels et matériels dotés de certaines formes d'intelligence.

Selon Grace Solomonov, la femme de **Ray Solomonov**, l'un des participants du Summer Camp de Darmouth, le terme d'IA a été choisi par John McCarthy parce qu'il était neutre par rapport aux sciences existantes comme la cybernétique[16]. Elle évoque aussi les sources de science-fiction qui inspirèrent son mari (*ci-dessous*).



Science fiction from 1953, 1954 and 1957.

L'IA est en fait une appellation créée par un chercheur afin de faire parler de son domaine et lui permettant d'éviter d'être assimilé à des disciplines voisines comme les mathématiques, les statistiques ou l'informatique. C'est une forme de déclaration d'indépendance d'une nouvelle discipline scientifique.

Plus de 60 ans plus tard, l'IA décrit aussi bien le champ du possible d'aujourd'hui dans ces domaines que la quête permanente et insatisfaite de l'incorporation des différents aspects de l'intelligence humaine dans des machines. L'IA a atteint l'âge de la retraite mais est encore adolescente et brouillonne.

---

[14] Pour la petite histoire, 1955 est aussi l'année de la naissance de Steve Jobs et Bill Gates.Tout un symbole !

[15] Le document historique est A proposal for the Darmouth summer research project of artificial intelligence, 31 août, 1955. Il prévoyait un budget de $13500, les deux tiers correspondant aux salaires des participants, situés entre $600 et $700 par mois, ce qui équivaudrait à environ $6000 d'aujourd'hui, charges comprises. La note a été envoyée à la Fondation Rockefeller le 2 septembre 1955. Le bilan de Solomonov n'était pas enthousiasmant : "*The research project wasn't very suggestive. The main things of value : 1) wrote and got report reproduced (very important) 2) Met some interesting people in this field 3) Got idea of how poor most thought in this field is 4) Some ideas: a) Search problem may be important. b) These guys may eventually invent a Turing Machine simply by working more and more interesting special problems*".

[16] Elle a rédigé en 2006 un document qui relate l'Histoire détaillée de ce Summer Camp : Ray Solomonoff and the Dartmouth Summer Research Project in Artificial Intelligence, 1956 (28 pages).

"théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence"

dans la pratique, pour l'instant : l'IA **simule**, **complète** et parfois **dépasse** largement des composantes **isolées** de l'intelligence humaine

aucune IA n'a pour l'instant les **capacités intégratives** de l'intelligence humaine

si on comprend la méthode, ce n'est plus de la **magie** comme dans le machine learning de base

ce n'est pas de l'IA et tant que le **test de Turing** n'est pas passé, l'IA n'existe pas !

"POUR MOI, L'IA N'EXISTE PAS."

VICE-PRÉSIDENT DU STRATEGY AND INNOVATION CENTER DE SAMSUNG ELECTRONICS ET CO-INVENTEUR DE SIRI, LUC JULIA RESTE TRÈS CRITIQUE, VOIRE SCEPTIQUE QUANT AUX SUPER POUVOIRS ATTRIBUÉS AUX NOUVELLES TECHNOLOGIES.

LUC JULIA
VP Innovation - Samsung Electronics

CBN .. Vous intervenez le 14 juin lors de l'IMM start-up contest sur le thème «l'AI n'existe pas !». Qu'entendez-vous par là ?

LUC JULIA.. L'IA pour moi n'existe pas, je l'appellerai plutôt *machine learning* ou connaissance artificielle. Il faut être clair sur les techniques et l'avancement de ce que l'on appelle IA : on a mis de la connaissance dans des machines qui sont devenues plutôt bonnes dans les domaines précis – jouer aux échecs, reconnaître un chat… – mais on ne leur a pas donné de l'intelligence.

CBN .. Que peut-on attendre de l'IA ?

LUC JULIA.. La techno en général et l'IA en particulier va simplifier la vie des gens. Nous travaillons par exemple sur l'IoT (*Internet of Things*, ndlr) pour que les objets connectés puissent travailler entre eux, faire de la prédiction et, par exemple, m'expédier une pièce détachée de machine à laver quelques jours avant que la panne ne survienne ! Dans l'éducation aussi la valeur de l'IA est indéniable en supprimant notamment certaines inhibitions qui freinent l'apprentissage (la présence d'un professeur ou d'autres élèves…).

CBN .. Vous avez créé Siri, vous vous montrez pourtant très critique à l'égard des assistants vocaux : pourquoi ?

LUC JULIA.. La reconnaissance de la parole générique ne peut pas fonctionner car les techniques n'ont pas évolué depuis 20 ans. On est plus rapide, plus précis dans des domaines particuliers mais vous n'arriverez jamais à un assistant comme dans *Her* (le film de Spike Jonze, sorti en 2013, ndlr). Nous n'y arriverons jamais car il est impossible de modéliser le monde dans sa totalité. Dans un échange simple et linéaire, tout ira bien, mais dans le cadre d'une conversation entre nous où l'on peut partir dans tous les sens, c'est terminé. Comme nous l'avons fait pour Siri, la machine compensera son incapacité par des artifices.

CBN .. Le fait de converser naturellement – d'égal à égal – avec un robot, comme le font déjà certains enfants, n'est-il pas dangereux ?

LUC JULIA.. Socialement, c'est très dangereux car une machine ne peut pas vous donner ce que peut donner un humain. Elle ne peut vous donner que ce qu'elle contient, un corpus de données… potentiellement biaisé. C'est un travail d'éducation.

CBN .. Quid des autres innovations : réalité virtuelle ou augmentée ?

LUC JULIA.. La première n'a, dans le domaine scientifique, aucun intérêt particulier (à part les jeux vidéo et parfois l'apprentissage dans les milieux difficilement accessibles) car elle crée un monde parallèle dans lequel nous sommes souvent mal à l'aise (*motion sickness*), et qui n'est pas le notre. Alors que la réalité augmentée a un vrai avenir, justement parce qu'elle vient s'appuyer sur la réalité pour nous donner des informations complémentaires, qui peuvent s'avérer très utile. Le bénéfice est clair. ■

CB NEWS . 34 . JUIN 2018

17

L'appellation la plus appropriée serait peut-être celle d'**intelligence humaine augmentée**, l'IA étant principalement destinée à démultiplier les capacités de l'Homme, comme tous les outils numériques jusqu'à présent, même si dans certains cas, l'IA peut effectivement se substituer aux travaux de l'Homme pour quelques tâches élémentaires comme, à relativement moyen terme, la conduite de véhicules[18].

Dans le domaine du raisonnement automatisé, l'IA est censée apporter une rationnalité dont l'Homme ne fait pas toujours preuve. Là encore, nous sommes dans l'ordre de la complémentarité.

L'intelligence artificielle incarne finalement la conquête d'un **Graal** distant, ayant été à l'origine, sur son chemin, d'un tas d'avancées technologiques relativement distinctes et plutôt complémentaires de l'intelligence humaine[19]. Celle-ci est encore unique dans la capacité à réagir avec discernement face à des situations nouvelles, à tirer profit de circonstances fortuites, à discerner le sens de messages ambigus ou contradictoires, à juger de l'importance relative de différents éléments d'une situation, à trouver des similitudes entre des situations malgré leurs différences, à établir des distinctions entre des situations malgré leurs similitudes, à synthétiser de nouveaux concepts malgré leurs différences ou à trouver de nouvelles idées[20].

Bases conceptuelles

Deux ans après le summer camp de Darmouth avait lieu le **Congrès de Middlesex** (1958) au Royaume Uni avec des contributions des principaux artisans du Congrès de Darmouth, Marvin Minsky et John MacCarthy ainsi qu'**Oliver Selfridge**, lui aussi présent à Darmouth. L'objet des publications associées était la modélisation et la mécanisation des mécanismes de la pensée en particulier avec des logiques heuristiques.

S'en suivirent des publications clés comme Some Methods of Artificial Intelligence and Heuristic Programming de Marvin Minsky qui jettait les bases théoriques de la programmation heuristique approfondie peu après dans Steps Toward Artificial Intelligence de Marvin Minksy.

[17] L'illustration de droite est une interview « Pour moi l'IA n'existe pas » de Luc Julia, parue dans CBNews en juin 2018.

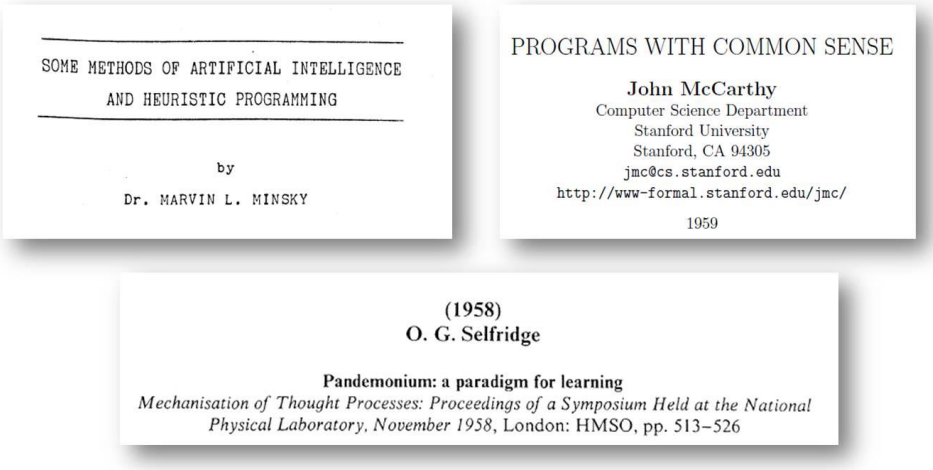
[18] Voir Intelligence artificielle : en finir avec les mirages de Vincent Champain, avril 2018 qui évoque cette notion d'intelligence augmentée.

[19] On pourrait dire qu'il en va de même des cancérologues dont le métier est de guérir le cancer et qui n'y arrivent pas forcément.

[20] La source de cette énumération est le cours d'intelligence artificielle d'Olivier Boisard.

congrès de Middlesex - UK - 1958

mechanization of thought processes



La même année, Pandemonium : a paradigm for learning d'**Oliver Selfridge**, jettait les bases des réseaux de neurones pour la reconnaissance des formes, puis Programming with common sense de John McCarthy, celle des systèmes experts. McCarthy est aussi connu pour être le créateur à la même époque du langage **LISP** qui servit pendant plusieurs décennies à développer des solutions logicielles d'IA travaillant en logique formelle et à base de règles.

Les années 1960 furent une période de recherche fondamentale importante, notamment au **MIT AI Lab**. Ces recherches étaient principalement financées par l'ARPA, l'agence de recherche du Pentagone créée en 1958, devenue la DARPA en 1972, l'équivalent de la DGA française, mais évidemment bien mieux financée avec un peu plus de $3B[21] de budget annuel actuellement.

La recherche sur l'IA était financée par les deniers publics, notamment aux USA et au Royaume-Uni. Encore aujourd'hui, une très grande partie des recherches les plus avancées sur l'IA aux USA sont financées par l'omniprésente DARPA et réalisées par des laboratoires de recherche d'Université ainsi que chez les GAFA, ainsi que par les agences liées au renseignement comme la CIA et la NSA. Ce qui peut alimenter au passage les craintes sur les applications futures de l'IA, notamment au moment hypothétique où elle atteindrait le stade de l'AGI (IA généraliste).

En **France**, on peut noter quelques dates clés avec la création de l'**INRIA** en 1967, à l'époque IRIA[22], dans le cadre du plan calcul visant à apporter une indépendance à la France dans les calculateurs nécessaires au développement de la dissuation nucléaire.

La création du langage **Prolog** par Alain Colmerauer[23] et Philippe Roussel à l'Université d'Aix-Marseille de Luminy date de 1972. La recherche française en IA ne semble cependant avoir véritablement décolé qu'à la fin des années 1970. Elle s'est plutôt spécialisée dans l'IA symbolique, un domaine dont on entend peu parler depuis le tsunami mondial du deep learning lancé à partir de 2006. Il recouvre diverses techniques de modélisation des connaissances et du raisonnement.

Prouesses de démonstrations de théorèmes

Les premiers travaux autour de l'IA portèrent sur la logique formelle et sur la démonstration automatique de théorèmes, surtout en géométrie.

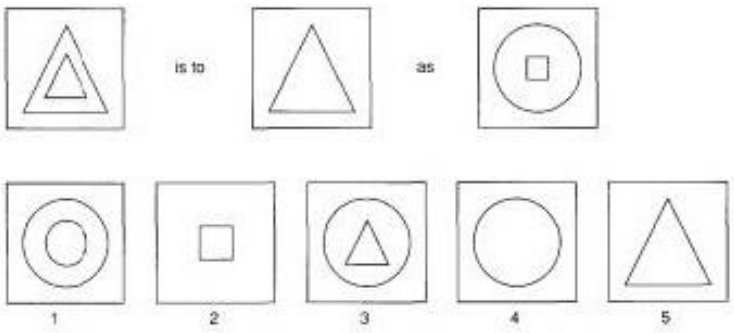
---

[21] J'utilise la nomenclature US pour les montants en dollars. $3B veut dire trois milliards de dollars. En Euros, j'utilise 3Md€. Lorsque je cite une startup, j'indique généralement entre parenthèses son année de création, le pays d'origine et la totalité des financements obtenus, récupérés le plus souvent sur Crunchbase.

[22] Le IA signifiait Informatique et Automatisme, pas Intelligence Artificielle !

[23] Alain Colmerauer est décédé en mai 2017.

Le premier en date est le **Geometry Theorem Prover** d'Herbert Gelernter de 1959, un logiciel de démonstration de théorèmes de géométrie fonctionnant en chaînage arrière - de la solution jusqu'au problème - sur un IBM 704 à lampes et à partir d'une base de 1000 règles. Cela relevait d'une combinatoire assez simple. C'était plutôt prometteur.



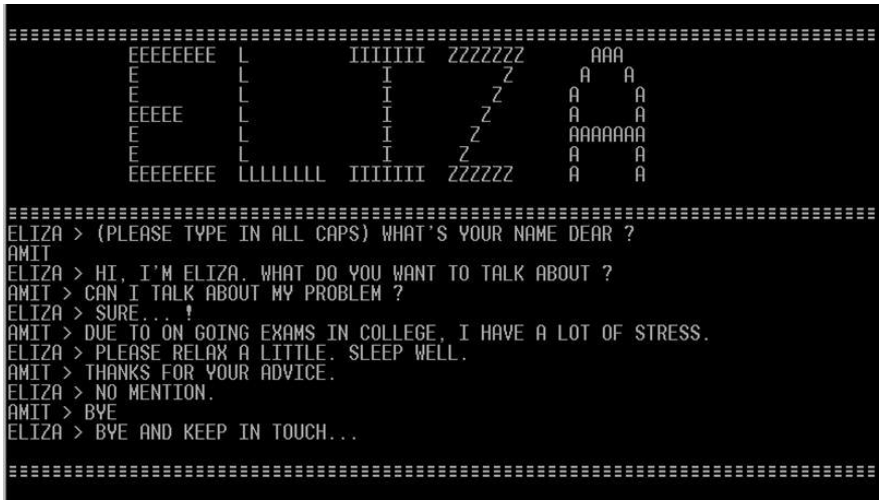
Suivirent le **General Problem Solver** d'Allen Newell et Herbert Simon en 1959, l'**Integration Problems Solver** de James Slagles en 1963, le **Geometric Analogy Problems** de Tom Evans en 1968, qui traitait les problèmes de géométrie qui sont intégrés dans les tests de quotient intellectuel (*ci-dessus*) et puis l'**Algebra Problems Solver** de Daniel Bobrow en 1967. Tout cela bien avant les débuts de la micro-informatique !

Les méthodes créées pour ces prouesses servirent plus tard de base aux techniques de moteurs de règles et de systèmes qui connurent leur heure de gloire pendant les années 1980.

Premiers chatbots

On vit aussi apparaître les ancêtres de catégories de solutions d'IA courantes aujourd'hui avec l'un des premiers chatbots, simulant un dialogue avec un psy, **ELIZA** entre 1964 et 1966.

Suivra **SHRDLU**, de Terry Winograd du MIT, l'un des premiers à comprendre le langage naturel en 1968 (vidéo).



Ces premiers chatbots tenaient le coup pendant des conversations avec quelques échanges mais ne passaient pas le test de Turing. Malgré tout, ils n'ont pas à rougir vis-à-vis de nombreux chatbots contemporains.

Surpromesses et premier hiver de l'IA

L'IA connut son premier "hiver" avec une réduction d'une bonne part de ses budgets de recherche à partir du début des années 1970, tant au Royaume-Uni qu'aux USA. En cause, les surpromesses des scientifiques du domaine et les lentes avancées associées.

Aux USA, le Congrès votait les **Mansfield Amendments** dans le Defense Procurement Act de 1969 puis 1973, du nom d'un sénateur démocrate, demandant à ce que la recherche financée par l'ARPA, l'ancêtre de la DARPA, ait des applications directes dans l'armée. En gros, cela coupait les financements de la recherche fondamentale civile, dont celle de l'IA et en particulier les travaux de **BBN Technologies**, un sous-traitant de l'ARPA situé à Cambridge dans le Massachusetts, créé par des anciens chercheurs du MIT et filiale de Raytheon depuis 2009. On doit à BBN la mise en œuvre du réseau ARPANET et l'invention de l'email en 1971 !

Ironie de l'Histoire, par un phénomène de vases communiquants, cela aurait contribué indirectement à la naissance de l'industrie de la micro-informatique au milieu des années 1970[24] !

Ce premier hiver de l'IA a duré jusqu'en 1980. Pendant ce trou d'air, nous avons eu entre autres la création du **Micral** en France (1973), de **Microsoft** (1975), d'**Apple** (1977), d'**Oracle** (1977), le Star de **Xerox** (1980 qui a ensuite inspiré Apple pour Lisa (1983) et le Macintosh (1984) puis les préparatifs du lancement de l'**IBM PC** (1980-1981).

Au Royaume Uni, l'hiver était la conséquence de la publication du **Rapport Lighthill** destiné à l'organisme public britannique **Science Research Council** – équivalent de notre Agence Nationale de la Recherche française d'aujourd'hui – qui remettait en cause le bien fondé des recherches de l'époque en robotique et en traitement du langage.

Une approche bien curieuse quand on sait que les technologies informatiques matérielles sous-jacentes n'étaient pas encore bien développées à cette époque[25]. C'est un bel exemple de manque de vision long terme des auteurs de ce rapport.

En cause dans le rapport Lighthill, des promesses trop optimistes des experts du secteur. Comme souvent, les prévisions peuvent être justes sur tout ou partie du fond mais à côté de la plaque sur leur timing[26].

Herbert Simon et **Allen Newell** prévoyaient en 1958 qu'en dix ans, un ordinateur deviendrait champion du monde d'échecs et qu'un autre serait capable de prouver un nouveau et important théorème mathématique. Trente ans d'erreur de timing pour la première prévision et autant pour la seconde sachant qu'elle est toujours largement en devenir pour être générique !

Herbert Simon (*ci-contre*) prévoyait– toujours en 1958 – qu'en 1978, les machines seraient capables de réaliser toutes les activités intellectuelles humaines. Et la loi de Moore n'existait pas encore puisqu'elle a été énoncée bien après cette prévision, en 1965 et observée entre les années 1970 et 2010. En 1967, **Marvin Minsky** pensait qu'en une génération, tous les problèmes liés à l'IA seraient résolus. Deux générations plus tard, on en discute encore. Il prévoyait aussi qu'au milieu des années 1970, les ordinateurs auraient l'intelligence d'un homme moyen.

Reste à savoir ce qu'est un homme moyen. Moyen vraiment moyen, ou juste moyen moyen ? Et combien de robots peuvent courir un marathon ?

Les retards étaient manifestes dans la traduction automatique et dans la reconnaissance de la parole. Notons qu'Herbert Simon a été récompensé en 1978 par le Prix Nobel d'économie, pour ses travaux sur les rationalités de la prise de décision, après avoir gagné la fameuse médaille de Turing en 1975. Il n'existe pas encore de prix Nobel de la prévision ! Il faudrait d'ailleurs plutôt les attribuer à des personnes déjà décédées pour valider leurs prévisions au long cours !

Ces prévisions trop ambitieuses ont existé de tous les temps. Leurs versions actualisées tournent autour de la singularité et du transhumanisme : l'ordinateur plus intelligent que l'homme entre 2030

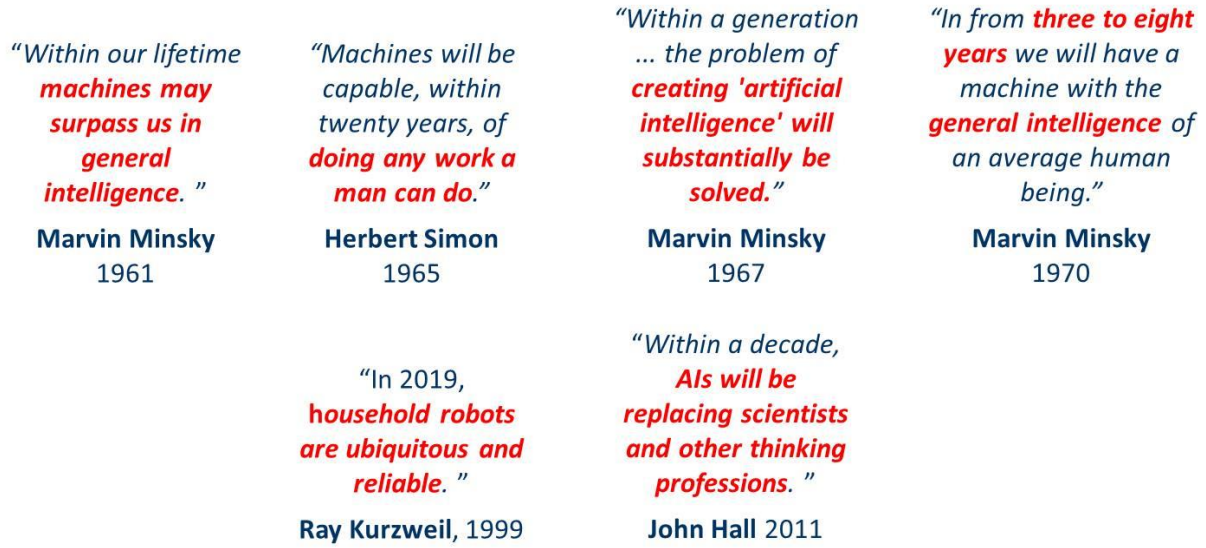
---

[24] Comme relaté dans la fiche Wikipedia sur la DARPA : https://en.wikipedia.org/wiki/DARPA.

[25] 1973 est l'année de l'apparition du premier micro-ordinateur de l'histoire, le français Micral de François Gernel et André Truong.

[26] Machine who think de Pamela McCorduck, 2004 (584 pages) en fait un inventaire intéressant en relatant les cinquantes premières années de l'Histoire de l'intelligence artificielle. Voir aussi cette timeline de History of artificial intelligence.

ou 2045 et l'immortalité ou une vie de 1000 ans pour les enfants qui viennent de naître ! Vous pouvez choisir, mais ne pourrez pas forcément valider vos propres hypothèses en temps et en heure !



Première renaissance avec les systèmes experts et nouvel hiver

Ce premier hiver a été suivi d'une période d'enthousiasme au début des années 1980 alimentée notamment par la vague des systèmes experts. Cet enthousiame a duré moins d'une décennie. Une nouvelle vague de désillusions s'en est suivie autour des années 1990. Notamment du fait de l'essouflement de la vague des systèmes experts et l'effondrement associé du marché des **ordinateurs dédiés au langage LISP**.

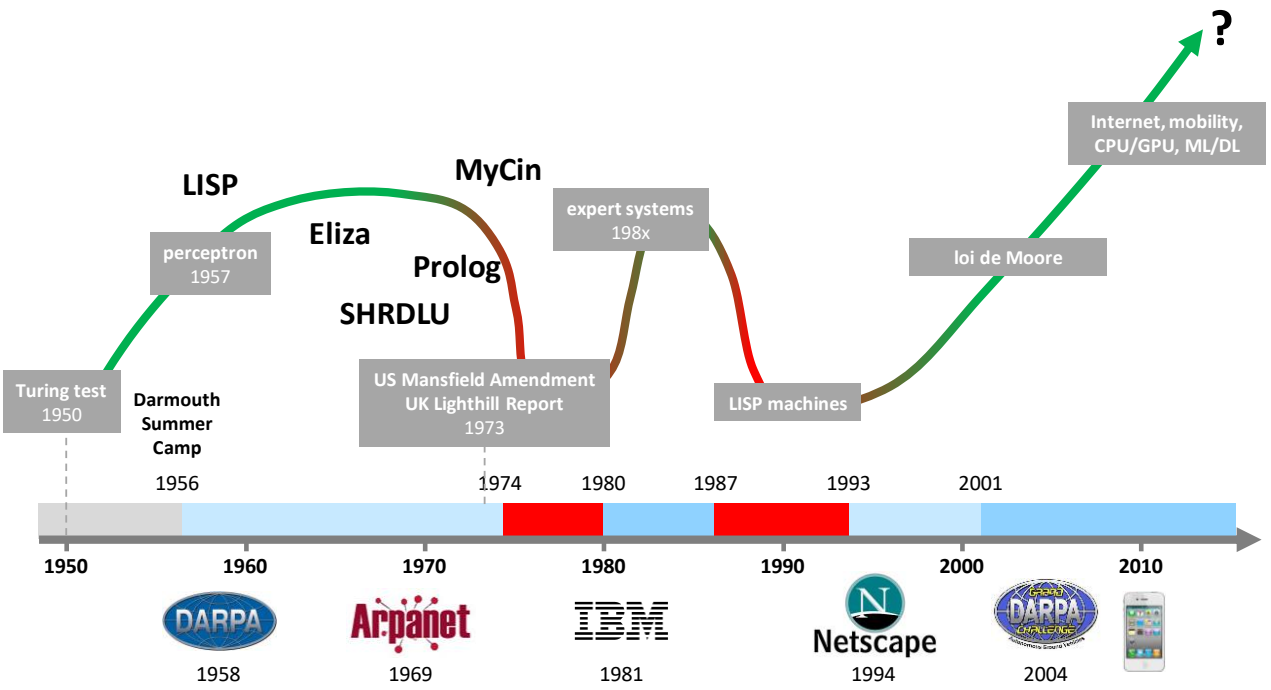
L'autre raison était que le matériel n'arrivait pas à suivre les besoins de l'IA, notamment pour traiter deux besoins clés : la reconnaissance de la parole et celle des images, très gourmandes en puissance de calcul, même avec les algorithmes de l'époque qui n'exploitaient pas encore les réseaux de neurones d'aujourd'hui. Il s'agissait des débuts de l'approche connexionniste (1980-1995).

Lors des années 1980 avaient été lancés divers *gosplans* d'ordinateurs "*de cinquième génération*" dédiés aux applications de l'IA. Cela a commencé avec celui du **MITI Japonais,** lancé en 1981 avec un budget de $1B (un milliard de dollars), puis avec le projet anglais **Alvey** doté de £350 million et enfin, avec le **Strategic Computing Initiative** de la DARPA. Tous ces projets ont capoté et ont été clôturés discrètement.

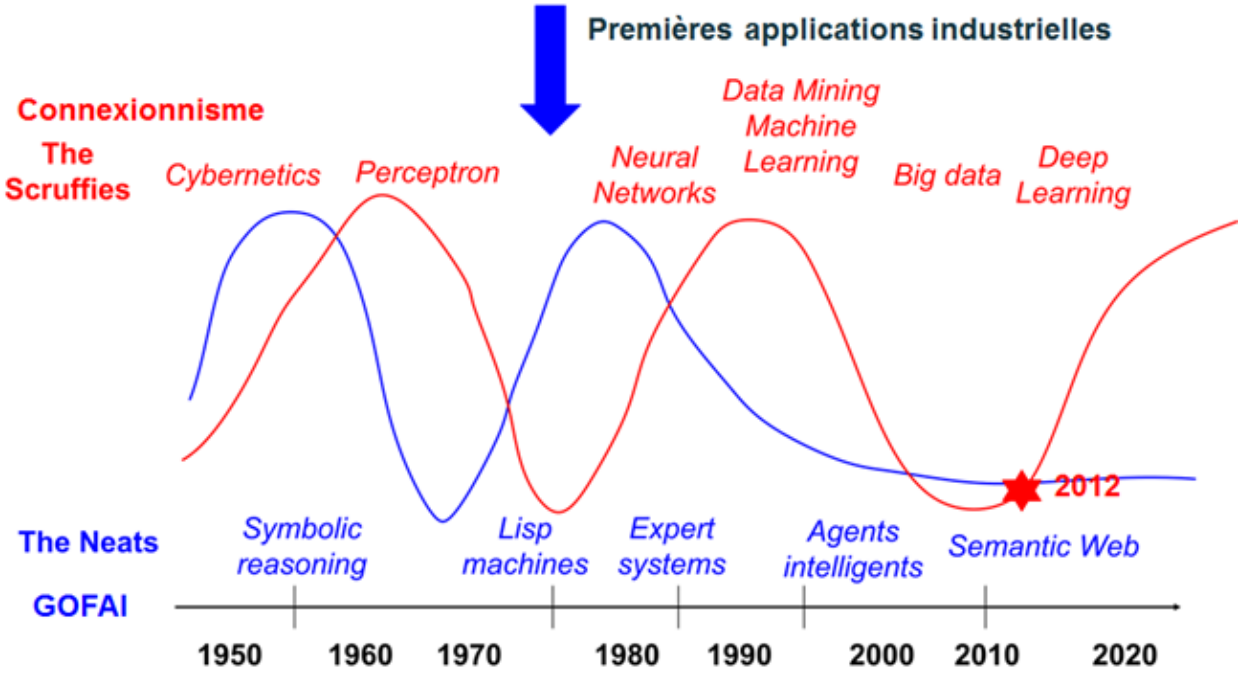
Le projet du MITI visait à faire avancer l'état de l'art côté matériel et logiciel. Les japonais cherchaient à traiter le langage naturel, à démontrer des théorèmes et même à gagner au jeu de Go. Le projet a probablement pâti d'une organisation trop traditionnelle, hiérarchique et centralisée.

Pendant les années 1990 et 2000 ont émergé de nombreux projets de **HPC** (high-performance computers), assez éloignés de l'IA et focalisés sur la puissance brute et les calculs de simulation par éléments finis. Ils étaient et sont encore utilisés pour de la simulation, notamment d'armes nucléaires, d'écoulements d'air sur les ailes d'avion ou pour faire des prévisions météorologiques. Les HPC de **Cray Computers** avaient été créés pour cela ! Cette société existe **toujours**. C'est l'une des rares survivantes des années 1970.

Le schéma *ci-dessous* de mon cru illustre ces hauts et ces bas vu de haut.



En voici, *ci-dessous* une variante, provenant de Françoise Soulié-Fogelman qui présente le décalage entre les étés et les hivers des IA symbolique et connexionniste. Les réseaux de neurones et le deep learning ont connu un premier essort au milieu des années 1990 (évoqué dans cette partie sur les débuts du deep learning) avant de s'essouffler dans les années 2000 puis de redécoler véritablement à partir de 2012.

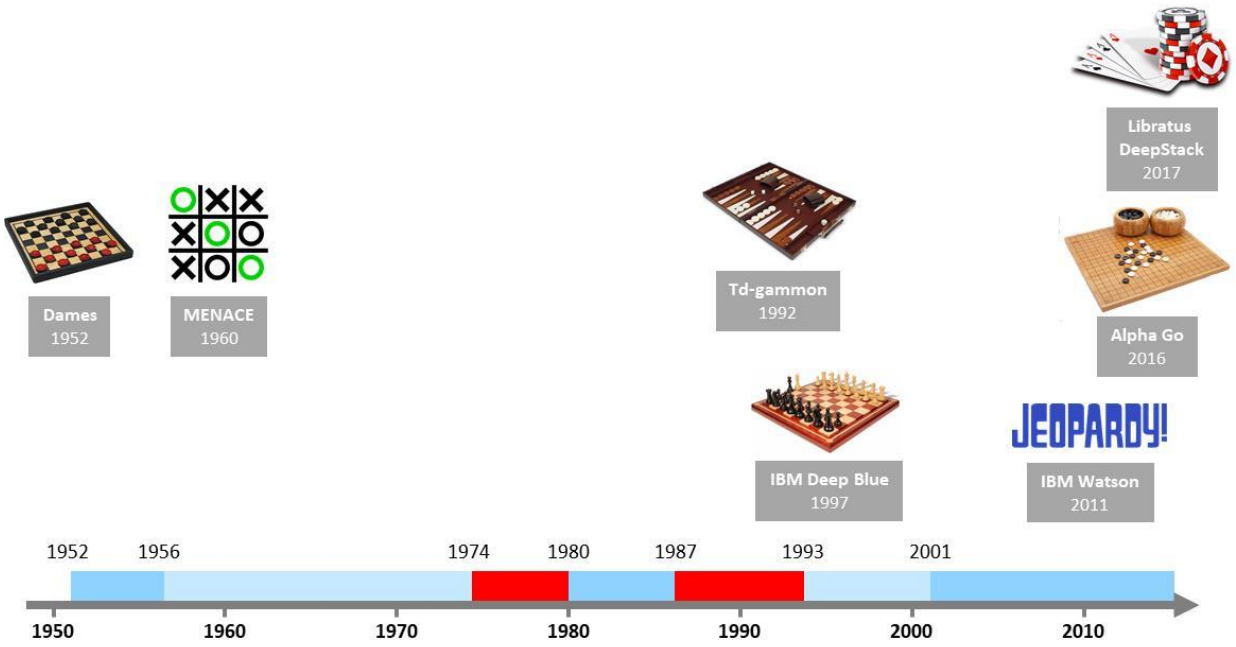


Dernière renaissance de l'IA

Depuis le début des années 2000, et surtout depuis 2012, l'IA a été relancée grâce à diverses évolutions majeures :

- Les **progrès théoriques et pratiques** constants dans le machine learning, les réseaux de neurones et le deep learning. Nous aurons l'occasion de tous les évoquer plus en détail dans la seconde partie dédiée aux algorithmes et logiciels de l'IA.

- L'augmentation de la **puissance du matériel** qui a permis de diversifier la mise en œuvre de nombreuses méthodes jusqu'alors inaccessibles. Et en particulier, l'usage du machine learning pouvant exploiter la puissance des machines autant côté calcul que stockage et puis, plus récemment, les réseaux neuronaux et le deep learning. Cette augmentation de puissance se poursuit inexorablement, malgré les limites actuelles de l'intégration des transistors dans les circuits intégrés. Une partie de ce document est aussi dédiée à cet aspect.



- L'atteinte de diverses **étapes marquantes** comme la victoire d'IBM Deep Blue contre Kasparov en 1997 puis d'IBM Watson dans Jeopardy en 2011. Enfin, début 2016, la victoire de Google DeepMind AlphaGo au jeu de Go contre le champion du monde. Les premiers jeux de société gagnés via l'IA étaient le jeu de dames (Checkers) et le tic-tac-toe dans les années 1950-1960. Il y avait eu près de 30 ans de calme plat dans le domaine des jeux de société. Depuis, deux IA ont aussi gagné au jeu de poker[27], Libratus et DeepStack ! Par rapport aux échecs ou au jeu de Go où le jeu est entièrement visible, la performance de ces IA tient au fait qu'elles agissent dans un environnement d'information incomplet et aussi, au fait qu'elles peuvent moduler l'aggressivité du jeu. Fin 2017, le logiciel AlphaGo de DeepMind était adapté pour jouer et gagner au jeu d'échecs et à sa variante japonaise du Shogi en apprenant par lui-même à jouer[28]. Enfin, en mai 2018, le laboratoire de recherche en IA de Facbebook (FAIR) publiait OpenGo, sa version open source d'une IA gagnant aussi au jeu de Go face à d'autres IA publiques, mais pas face à AlphaGo Zero qui n'est pas en open source[29]. Tous ces succès récents s'appuient sur de l'IA connexionniste et du deep learning.

[27] Voir Artificial intelligence goes deep to beat humans at poker, mars 2017. La description technique de DeepStack, créé par des chercheurs canadiens et tchèques, est dans DeepStack: Expert-Level Artificial Intelligence in No-Limit Poker. Celle de Libratus, créé par Tuomas Sandholm et NoamBrown, de l'Université Carnegie Mellon de Pittsburgh est dans Libratus: The Superhuman AI for No-Limit Poker et pour la vulgarisation, dans cet article de Wired, Inside Libratus : the poker AI that out-bluffed the best humans. Dans les deux cas, il s'agissait de parties 1 contre 1. DeepStack et Libratus sont bien décrits dans cette présentation technique : Poker AI: Equilibrium, Online Resolving, Deep Learning and Reinforcement Learning de Nikolai Yakovenko (Nvidia), avril 2017. La prochaine étape sera d'intégrer à ces IA des capteurs sensoriels permettant de détecter les émotions des joueurs humains. A distance et avec une caméra, on peut détecter de fines variations dans les expressions et même la variation du pouls !

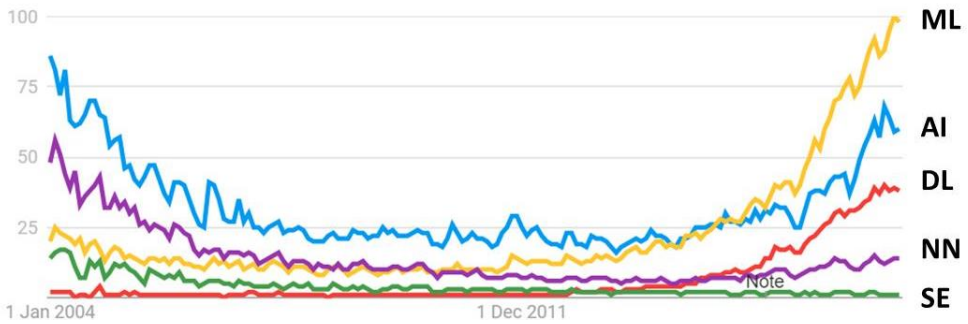
[28] Voir Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, décembre 2017 (19 pages).

[29] Voir Facebook Open Sources ELF OpenGo de Yuandong Tian et Larry Zitnick, mai 2018.

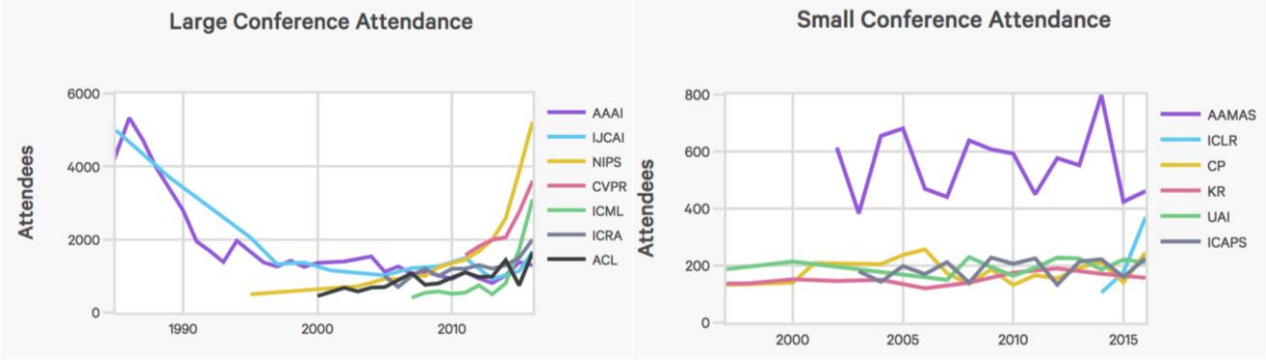
- L'**Internet grand public** qui a créé de nouveaux besoins comme les moteurs de recherche et aussi permis le déploiement d'architectures massivement distribuées. L'Internet a aussi permi l'émergence de méthodes de travail collaboratives dans la recherche et les développements de logiciels, en particulier dans l'open source. Il a aussi fait émerger les fameux GAFA, ces acteurs dominants du Web grand public qui sont aussi très actifs dans l'IA. Le phénomène est en train de se démultiplier en Chine avec les fameux BATX et leur capacité à accumuler des données sans trop se soucier des règles sur la vie privée comme le RGPD européen.
- La disponibilité de très **gros volumes de données**, via les usages de l'Internet et des mobiles, des objets connectés ou de la génomique, exploitables par différentes méthodes de machine learning et de deep learning. Le travail des chercheurs et développeurs est facilité par la publication de jeux de données ouverts (open data) pour le machine et le deep learning avec de nombreuses bases d'images et de textes disponibles pour réaliser des benchmarks. C'est le cas de la base ImageNet, de la base décriture manuscrite MNIST et de la base linguistique WordNet (en anglais). Ces bases sont généralement d'origine américaine et proviennent le plus souvent de grandes Universités.
- La culture de l'**open source** qui domine les outils de développement de solutions d'IA et les jeux de données suscités. Les chercheurs doivent publier des exemples de codes sources pour illustrer leurs méthodes, sur Github qui peuvent alors être reproduits et vérifiés par la communauté des chercheurs et développeurs. Ce processus permet une diffusion rapide des nouveautés algorithmiques, particulièrement autour des réseaux de neurones et du deep learning. C'est par exemple le cas des réseaux de neurones génératifs qui colorient automatiquement les photos chez Google ou ceux qui font de l'upscaling de vidéos chez Samsung.
- L'**appel d'air** généré par la robotique, la conquête spatiale (Curiosity, Philae…), les véhicules à conduite assistée ou autonome, la sécurité informatique, ainsi que la lutte contre la fraude et les scams. Sans compter l'objectif des japonais de s'occuper de leurs seniors avec des robots faute d'accepter politiquement une immigration de travail.
- Les **nombreuses applications commerciales** de l'IA mêlant le machine learning, les objets connectés, la mobilité et le big data. Avec des attentes fortes dans le marketing, le e-commerce, la finance et le vaste secteur de la santé.

Comme les usages de l'IA sont bien concrets et qu'ils touchent presque toutes les industries, on peut affirmer sans trop risquer de se tromper que la tendance est solide.

Pour en avoir le cœur net sur les tendances de l'IA, j'ai fait une petite comparaison des principaux termes (machine learning, artificial intelligence, deep learning, neural networks, expert systems) pour voir l'évolution de leur recherche dans Google. Cela donne le schéma suivant qui illustre un curieux phénomène : des recherches abondantes vers 2004, un trou autour de 2011 et une renaissance depuis. Pourtant, l'appellation d'intelligence artificielle était presque un gros mot il y a 10 ans. A l'origine, un biais statistique classique, parmi d'autres : la typologie et la masse des utilisateurs Internet n'était pas la même en 2004 qu'elle ne l'est en 2018 ! Un bon exemple du besoin de recul face à n'importe quelle courbe temporelle. L'explosion du deep learning date de 2012 et de la victoire de l'équipe de Geoff Hinton au concours de reconnaissance d'images ImageNet.



Une autre manière d'évaluer le timing des hivers est d'évaluer la participation aux grandes conférences de chercheurs en IA[30]. On peut observer un trou autour de 2000 sur la participation aux grandes conférences de l'IA et une explosion, notamment de NIPS, à partir de 2012. Par contre, l'audience des conférences très spécialisées est restée très stable.



Une des manières d'évaluer l'évolution du champ scientifique de l'IA est de se se plonger dans l'ouvrage de référence **Artificial Intelligence A Modern Approach**, de Stuart Russell et Peter Norvig. Sa dernière édition de 2009 fait la bagatelle de 1152 pages et a été longtemps l'un des B-A-BA pour les étudiants en informatique et en IA de premier cycle[31]. Elle donne la part belle aux agents, aux techniques de recherches apparentées à la force brute utilisée notamment dans les jeux de société, à la programmation par contraintes, au raisonnement avec les agents logiques, la logique du premier ordre, la planification, la représentation des connaissances et les méthodes probabilistes et bayésiennes. Les bases du machine learning y sont présentes mais pas encore le deep learning.

Connexionisme et symbolisme

Comme tout domaine scientifique complexe, l'IA n'a jamais été un terrain d'unanimité et cela risque de perdurer. Diverses écoles de pensée se disputent sur les approches à adopter[32].

On a vu au départ s'opposer :

- Les partisans du **connexionnisme** – utilisant le principe du machine learning, du biomimétisme, des réseaux de neurones et de l'apprentissage, les réseaux de neurones étant pour l'instant surtout utilisé pour les sens artificiels (vision, parole). C'est le raisonnement inductif qui réalise des prévisions et des généralisations à partir d'observations. Ces méthodes sont le plus souvent probabilistes.
- Les partisans du **symbolisme** qui préfèrent utiliser des concepts de plus haut niveau sans chercher à les résoudre via des procédés de biomimétisme. L'IA symbolique modélise le raisonnement logique et représente les connaissances avec des objets et des symboles formels les associant entre eux (appartient à, fait partie de, est équivalent à, …) via d'éventuelles contraintes. C'est un raisonnement déductif qui s'appuie sur la logique reposant sur des faits et règles connus. On l'utilise par exemple pour résoudre des problèmes de mathématiques mais aussi des problèmes d'optimisation divers.

J'illustre ces deux courants avec la manière dont on apprend aux enfants en bas âge à traverser la rue. La méthode connexionniste et inductive consisterait à les laisser y aller au hasard en espérant qu'ils apprennent de leurs erreurs et ne se fassent pas écraser et que les adultes découvrent par l'observation les erreurs des enfants ou que les enfants apprennent par eux-mêmes.

---

[30] Voir Artificial Index Report 2017 (101 pages).

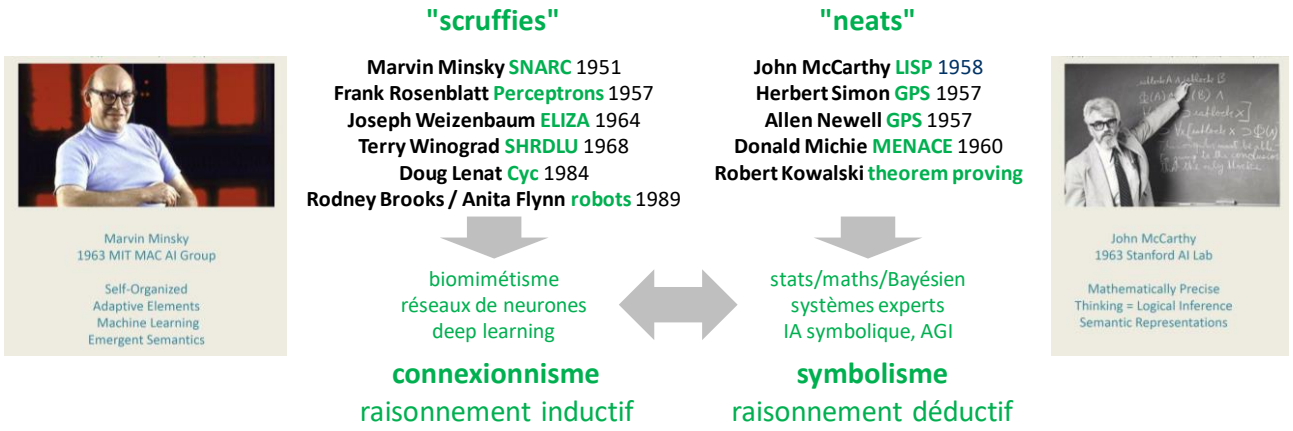
[31] Le Russell et Norvig 2010 est téléchargeable gratuitement ici. Ses auteurs prévoient de sortir une nouvelle édition en 2019 qui couvrira notamment les avancées du deep learning.

[32] Voir l'excellent La revanche des neurones - L'invention des machines inductives et la controverse de l'intelligence artificielle par Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières, novembre 2018 (38 pages) qui relate l'histoire de cette rivalité.

Ce serait du connexionnisme avec apprentissage par essais-erreurs ou par renforcement.

La méthode symbolique tire parti du savoir accumulé et des règles connues pour traverser la rue. On transmet aux enfants ces règles consistant à regarder des deux côtés de la route, à traverser sur les passages piétons, à respecter les feux et, éventuellement à traverser vite et à tenir la main de l'adulte accompagnant. Cela permet de gagner du temps et de sauver des vies !

Tout cela pour dire que si on connaît déjà les règles pour résoudre un problème ou faire une prévision, il vaut mieux les appliquer directement. On utilise la logique connexionniste lorsque l'on ne peut pas modéliser un système complexe avec des règles établies ou bien, lorsque ces règles changent très souvent et rapidement, comme dans le cas de la fraude, affectée par des fraudeurs toujours très inventifs.



Cette dichotomie était incarnée par la **joute intellectuelle entre "neats" et "scruffies"**, les premiers, notamment John McCarthy (Stanford), considérant que les solutions aux problèmes devraient être élégantes et carrées, et les seconds, notamment Marvin Minsky (MIT) que l'intelligence fonctionne de manière plus empirique et pas seulement par le biais de la logique[33]. Comme si il y avait un écart entre la côté Est et la côte Ouest !

Le schéma *ci-dessus* cite les protagnonistes les plus connus des années 1950 à 1970. On pourrait ajouter évidemment les grands chercheurs de l'IA connexionniste comme David Rumelhart, Geoff Hinton et Yann LeCun.

Ces débats ont leur équivalent dans les sciences cognitives, dans l'identification de l'inné et de l'acquis pour l'apprentissage des langues. **Burrhus Frederic Skinner** est à l'origine du comportementalisme linguistique qui décrit le conditionnement opérant dans l'apprentissage des langues. **Noam Chomsky** avait remis en cause cette approche en mettant en avant l'inné, une sorte de préconditionnement du cerveau des enfants avant leur naissance qui leur permet d'apprendre facilement les langues.

En gros, le fonctionnement de l'intelligence humaine est toujours l'objet de désaccords scientifiques ! On continue d'ailleurs à en découvrir toujours plus sur la neurobiologie et le fonctionnement du cerveau.

D'autres débats ont cours entre les langages de programmation déclaratifs et les moteurs d'inférences utilisant des bases de règles. Sont arrivées ensuite les méthodes statistiques s'appuyant notamment sur les réseaux bayésiens, les modèles de Markov et les techniques d'optimisation.

Après une dominance des méthodes mathématiques et procédurales, ce sont les réseaux de neurones et l'apprentissage profond les utilisant qui ont pris le dessus au début des années 1990 puis à partir de 2012, en particulier pour la vision artificielle, la reconnaissance et le traitement du langage.

[33] Voir Marvin Minsky said "Probabilistic models are dead ends." What approach are researchers taking to solve the problem of Artificial General Intelligence?, 2016 et Marvin Minsky: AI "has been brain dead since the 1970s", 2003.

La technique la plus remarquable étant celle des réseaux de neurones convolutionnels, créée par le français **Yann LeCun** et largement améliorée depuis par nombre de chercheurs, faisant elle-même suite aux TDNN (time delay neural networks) d'**Alexandre Waibel**[34]. Nous y reviendrons.

Pedro Domingos, l'auteur de « The Master algorithm », décompte en fait cinq grands courants dans l'IA en plus du symbolisme et du connexionnisme (*ci-dessous*).

Tribe	Origins	Problem	Master Algorithm
Symbolists	Logic, philosophy	Knowledge composition	Inverse deduction
Connectionists	Neuroscience	Credit assignment	Backpropagation
Evolutionaries	Evolutionary biology	Structure discovery	Genetic programming
Bayesians	Statistics	Uncertainty	Probabilistic inference
Analogizers	Psychology	Similarity	Kernel machines

Il faut ajouter celui des évolutionnistes avec les algorithmes génétiques (dont nous reparlerons), celui des bayésiens avec une vision probabiliste des choses et celui des analogistes et leurs algorithmes de clustering. Dans de nombreux cas, ces approches sont combinées pour générer des solutions optimales.

D'un point de vue historique, la vague symbolique a dominé l'IA entre ses débuts et les années 1980. Nous sommes actuellement en pleine vague connexionniste depuis l'essor du deep learning à partir de 2012.

Ce phénomène de vases communicants a eu un impact sur la capacité des IA à raisonner. On n'a pas fait de grands progrès dans ce domaine ces deux dernières décennies. En pratique, de nombreux chercheurs ambitionnent de fusionner les approches symboliques et connexionnistes pour gérer du raisonnement automatique[35].

Définitions et segmentations

L'IA est un ensemble de techniques permettant de résoudre des problèmes complexes en s'inspirant de mécanismes cognitifs humains, agissant de manière rationnelle en fonction de faits, données et expériences, et capables d'atteindre de manière optimale un ou plusieurs objectifs donnés.

La rationalité n'est pas l'omniscience mais la capacité à agir en fonction des informations disponibles, y compris celles qui sont ambigües. Cette rationalité est habituellement limitée par notre volonté, le poids émotionnel de notre cerveau limbique et notre capacité d'optimisation.

A haut niveau, on peut découper l'IA en trois grands domaines, dont deux que nous avons déjà décrits précédemment :

- Le **symbolisme** qui se focalise sur la pensée abstraite et la gestion des symboles, l'algorithmique et la logique. Le symbolisme modélise notamment les concepts sous la forme d'objets reliés entre eux par des prédicats logiques (appartient à, etc). C'est une approche « macro » de résolution de problèmes. C'est dans cette catégorie que l'on peut ranger les systèmes experts et moteurs de règles qui les font fonctionner, et dans une certaine mesure, le web sémantique.
- Le **connexionnisme** qui se focalise sur la perception, dont la vision, la reconnaissance des formes et s'appuie notamment sur les réseaux neuronaux artificiels qui reproduisent à petite échelle et de manière approximative le fonctionnement générique du cerveau. C'est une vision

[34] Les chercheurs français ont aussi fait avancer le domaine. Voir par exemple : Experiments with Time Delay Networks and Dynamic Time Warping for Speaker Independent Isolated Digits Recognition, de Léon Bottou, Françoise Soulié, Pascal Blanche, Jean-Sylvain Lienard, 1989 (4 pages) qui décrit une méthode de reconnaissance vocale de chiffre indépendante du locuteur.

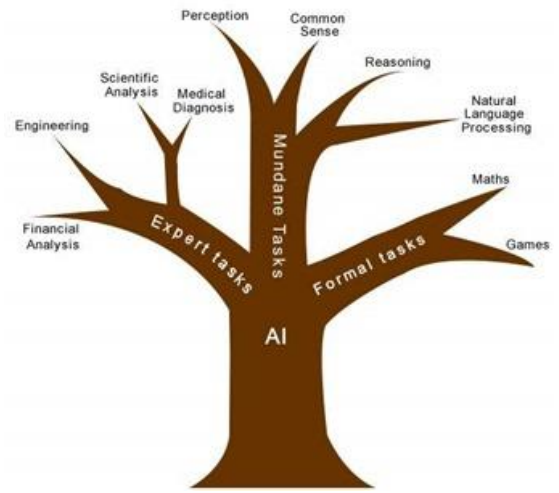
[35] Voir Que devient l'IA symbolique, que j'ai publié en avril 2018.

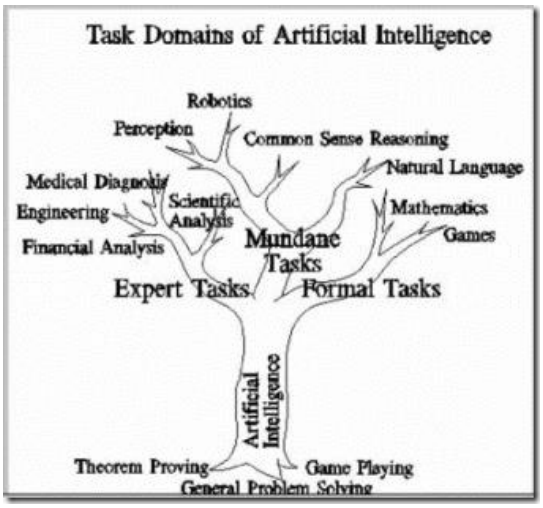
« micro » et probabiliste de la résolution des problèmes. C'est ici que l'on peut ranger le deep learning utilisé dans la vision artificielle ou le traitement de la parole. Cette IA est aussi associable aux méthodes stochastiques et heuristiques du machine learning.

- Le **comportementalisme** qui s'intéresse aux pensées subjectives de la perception. C'est dans ce dernier domaine que l'on peut intégrer l'informatique affective (ou affective computing) qui étudie les moyens de reconnaître, exprimer, synthétiser et modéliser les émotions humaines. C'est une capacité qu'IBM Watson est censé apporter au robot Pepper de Softbank Robotics (ex Aldebaran).

Reprenant plus ou moins ce découpage, cette autre segmentation sous forme d'arbre comprend trois grandes branches : l'une pour les **tâches d'expertise**, la seconde pour les **tâches courantes** (perception, sens commun, raisonnement, langage) et la troisième pour les **tâches formelles** (mathématiques, jeux).

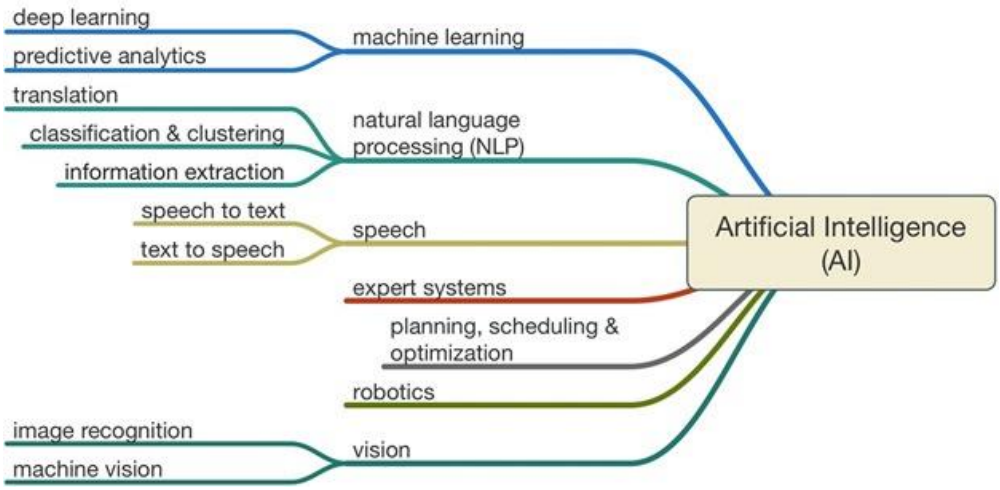
Cette autre segmentation très utilisée comprend le machine learning, le deep learning, le traitement du langage, les systèmes experts, la robotique et la vision.



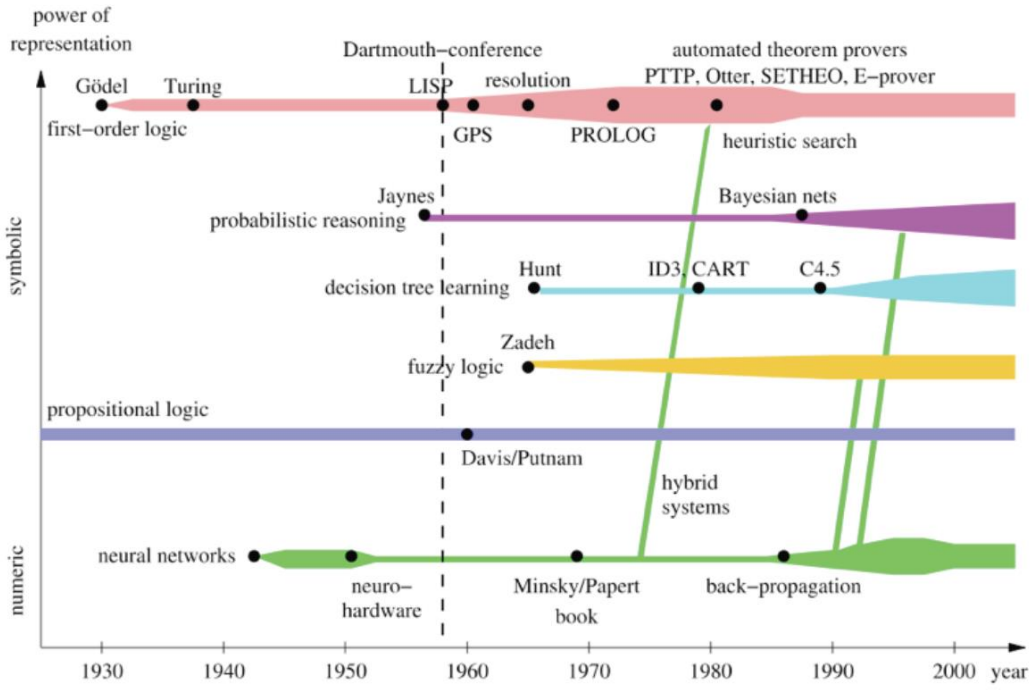


Comme de nombreuses tentatives de segmentation du champ de l'IA, elle place curieusement au même niveau des outils génériques comme le machine learning et le deep learning et ses propres applications comme la vision artificielle ou le traitement du langage.

La robotique intègre de son côté tous les autres champs du schéma plus quelques autres qui lui sont spécifiques comme les capteurs, les matériaux, la mécanique, les moteurs électriques et autres batteries.



Cette autre segmentation illustre l'évolution des grandes techniques de l'IA symbolique (en haut) et connexionniste (en bas) lors des 50 premières années de son Histoire. Bref, il faudrait inventer l'IA qui segmente convenablement l'IA !



Source: Ertel: Introduction to Artificial Intelligence, Springer, 2011.

Le rapport **France IA** publié en mars 2017[36] par le gouvernement français propose pour sa part une segmentation plus fouillée, compilant les principaux travaux de recherche du domaine en France. Mais cela reste encore touffu.

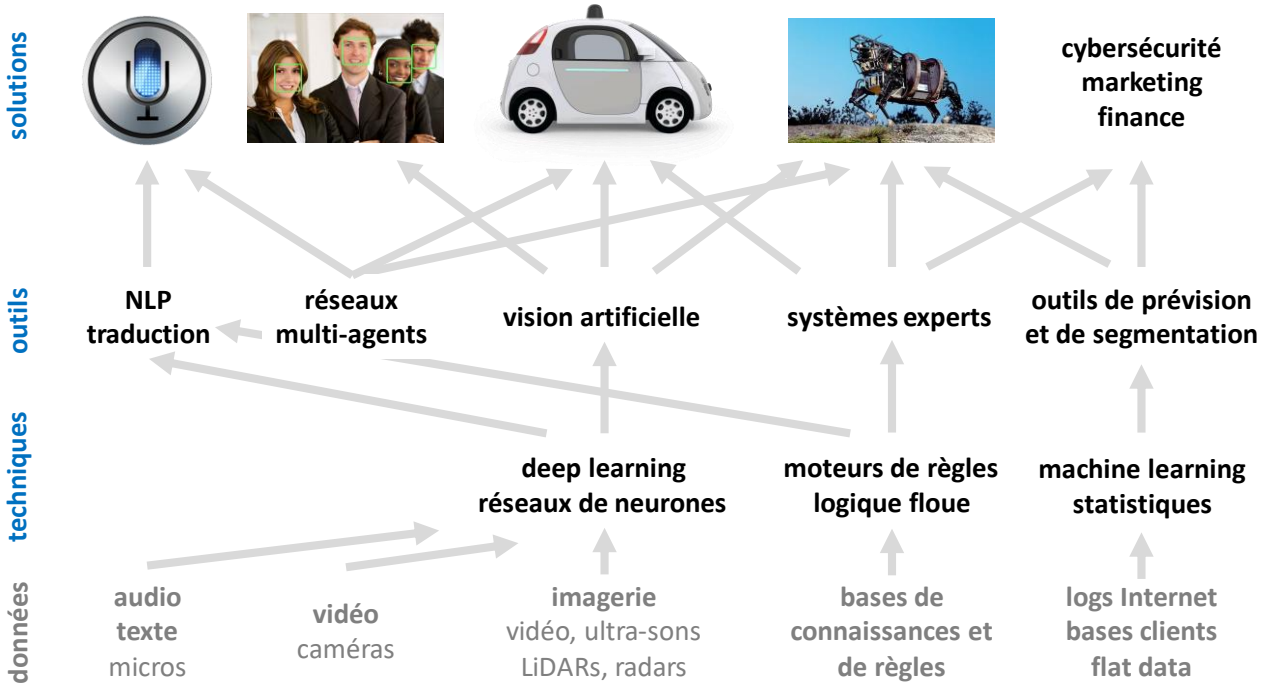
IA et SHS	Représentation des connaissances	apprentissage automatique	traitement du langage naturel	traitement des signaux	robotique*	neurosciences, sciences cognitives	algorithmique de l'IA	aide à la décision	systèmes multi-agents	interaction avec l'humain
Ethique	Bases de connaissances	Apprentissage supervisé / non-supervisé / séquentiel et par renforcement	Analyse syntaxique Lexiques Discours	Parole	Conception	Compréhension	Programmation logique et ASP		Coordination Multi-Agents (Planification multi-agents, Décision multi-agents)	Interaction avancée, apprentissage humain (EIAH)
Droit	Extraction et nettoyage de connaissances	Optimisation	(Interaction, Connaissances et Langage Naturel)	Vision	Perception	et stimulation du cerveau et du système nerveux	Déduction, preuve		Résolution Distribuée de Problèmes	
Economie	Inférence	Méthodes bayésiennes	Reconnaissance de la parole et traduction automatique	Reconnaissance d'objets	Décision	Sciences cognitives	Théories SAT		Apprentissage multi-agents	
Sociologie	Web sémantique	Réseaux de neurones ou neuronaux		Reconnaissance d'activités	Action		Raisonnement causal, temporel, incertain		Ingénierie Multi-Agents (Langages, plateformes, méthodologies)	
Humanités numériques	Ontologies	Méthodes à noyau		Recherche dans des banques d'images et de vidéos	Interactions avec les robots		Programmation par contraintes		Simulation Multi-Agents (intéresse aussi les SHS)	
		Apprentissage profond		Reconstruction 3D et spatio-temporelle	Flottes de robots		Recherche heuristique			
		Fouille de données		Suivi d'objets et analyse des mouvements	Apprentissage des robots		Planification et ordonnancement			
		Analyse de données massives		Localisation d'objets	Cognition pour la robotique et les systèmes					
				Asservissement visuel						

Enfin, voici ma propre proposition de segmentation qui relie entre eux quatre domaines de manière plus hiérarchique :

- Les **solutions** : que l'on va directement utiliser dans les entreprises ou chez les particuliers avec les chatbots, les véhicules autonomes, les robots, les systèmes de recommandation, les outils de segmentation client, le marketing prédictif ou les solutions de cybersécurité.

[36] Ici : Stratégie France I.A. : Pour le développement des technologies d'intelligence artificielle.

- Les **outils** : qui aident à créer ces solutions, comme la vision artificielle, la reconnaissance de la parole, la traduction automatique, les systèmes experts, les outils de prévision ou de segmentation automatiques. J'y ai ajouté les réseaux multi-agents qui coordonnent l'action des différents outils d'une solution d'IA.
- Les **techniques** : sur lesquelles sont construits ces outils, avec les méthodes de machine learning, les réseaux de neurones, les nombreuses méthodes de deep learning et les moteurs de règles.
- Les **données** : les sources de données correspondantes et les **capteurs** associés qui jouent un rôle indispensable, notamment pour les approches connexionnistes, le machine learning et le deep learning.



Cela rappelle que les solutions à base d'IA sont des assemblages de diverses briques logicielles et matérielles selon les besoins. Ces briques sont des plus nombreuses. A tel point que leur intégration est un enjeu technique et métier de taille, peut-être le plus complexe à relever[37].

Quand une startup indique qu'elle a créé « une IA » pour faire ceci ou cela, cela signifie qu'elle a assemblé des techniques, paramétré des outils, en général assez standards, pour exploiter des données, et les a appliqués pour créer une solution. L'originalité est rarement dans la création des briques mais plutôt dans leur sélection, leur combinaison, leur assemblage et le problème métier traité !

Limites de l'IA

Je vous livre ici le fruit d'une réflexion de quelques mois permettant de faire le tri de différentes méthodes qui relèvent de l'IA et de celles qui n'en sont pas. Le schéma *ci-dessous* illustre cela.

Les deux premières colonnes positionnent l'IA connexionniste et l'IA symbolique. La première extrait des règles de données observées et généralement taggées par l'Homme. Elle permet d'identifier

[37] Aymeric Poulain Maybant évoque sa thèse de doctorat dans L'Intelligence Artificielle en questions - Une spécialité qui se cherche en 2014. Il évoque l'hybridation en sciences cognitives qui date de 2005 et décrit très bien cet enjeu. L'IA intégrative est un des principaux facteurs de développement du secteur. On le retrouve dans l'association de nombreuses techniques dans les solutions d'IA comme le couplage de réseaux neuronaux et d'approches statistiques plus simples, notamment dans la reconnaissance de la parole.

de manière empirique des règles, par exemple sur le comportement de clients. Ces règles peuvent à leur tour alimenter une IA symbolique qui exploite des faits et règles connus et formalisés pour résoudre des problèmes de logique. En quelque sorte, le machine learning est une brique d'alimentation du raisonnement automatique[38].

La troisième colonne positionne l'interaction entre l'IA et le monde réel. C'est ce qui lui permet de faire de l'apprentissage par renforcement.

Elle vise à évaluer la réaction du monde réel à des actions externes. C'est le cas d'un chatbot qui évalue la qualité de son dialogue et de ses réponses. C'est celui du robot qui apprend par tâtonnement à maîtriser ses gestes, comme un enfant en bas âge qui apprend à marcher. Hors IA, c'est aussi le champ de l'expérimentation biologique en boîtes de Petri. Ces interactions s'appuient sur des mécaniques tellement complexes qu'elles ne sont pas simulables « in silico ». L'expérimentation avec le monde réel permet d'en extraire des règles qui peuvent à leur tour également alimenter de l'IA symbolique. Par exemple : telle protéine agit de telle manière sur tel type de cellule à telle température.

La dernière colonne positionne les situations et problèmes qui sont simulables ou partiellement simulables « in silico », dans des ordinateurs, sans passer par le monde réel et l'expérimentation de la colonne précédente.

On y trouve deux catégories de problèmes : ceux qui s'expriment avec des règles simples et discrètes et dont la simulation relève souvent de l'élagage d'un arbre de décision. C'est le cas des jeux de société comme le jeu d'échecs et celui de Go. Et puis celui de la physique et du monde biologique qui s'appuie sur des règles complexes, des équations différentielles ou linéaires, continues, et dont la simulation commence à être possible mais est difficile. Cela concerne par exemple la simulation du repliement des protéines dans les cellules, un des problèmes les plus complexes à résoudre qui soit pour de grandes protéines. C'est l'un des champs d'application du calcul quantique. Lorsque les règles sont discrètes comme avec le jeu de Go, la simulation s'effectue avec des réseaux de neurones et fait donc partie du champ de l'IA. Lorsque la simulation relève de règles continues et mathématiques, et qu'on les modélise avec du calcul quantique, on sort du champ de l'IA.

Il en va de même lorsque l'on réalise de la simulation de systèmes complexes par approximation, comme avec l'usage de la méthode des éléments finis pour des simulations d'écoulement d'air sur les ailes d'avions ou pour des simulations météo. Dans ces cas-là, on alimente les simulations avec des états réels mesurés, comme les images satellites et données de terrain pour la météo. On commence à parler de jumeau numérique pour représenter un système et effectuer des simulations dessus au lieu de faire les simulations dans le monde réel, comme pour les crash tests de véhicules ou la simulation de vol d'un aéronef.

Cette dernière colonne correspond aux limites de l'IA et des techniques d'origine humaine. Elles sont liées aux théories de la complexité qui régissent la nature des problèmes que l'on peut résoudre ou pas avec des calculs[39].

Dans la seconde et la quatrième colonne, des problèmes de logique ou de simulation sont tellement complexes qu'ils sont indécidables. Ils ne peuvent pas avoir de solution. C'est pour cela que l'IA ne prévoit généralement pas les « cygnes noirs » et que la société ne peut pas se mettre facilement en équation[40]. On en est alors réduit à faire du réductionnisme, en réduisant par grandes approximations des problèmes complexes à un ensemble de problèmes plus simples. Et on est obligé de faire des expériences.

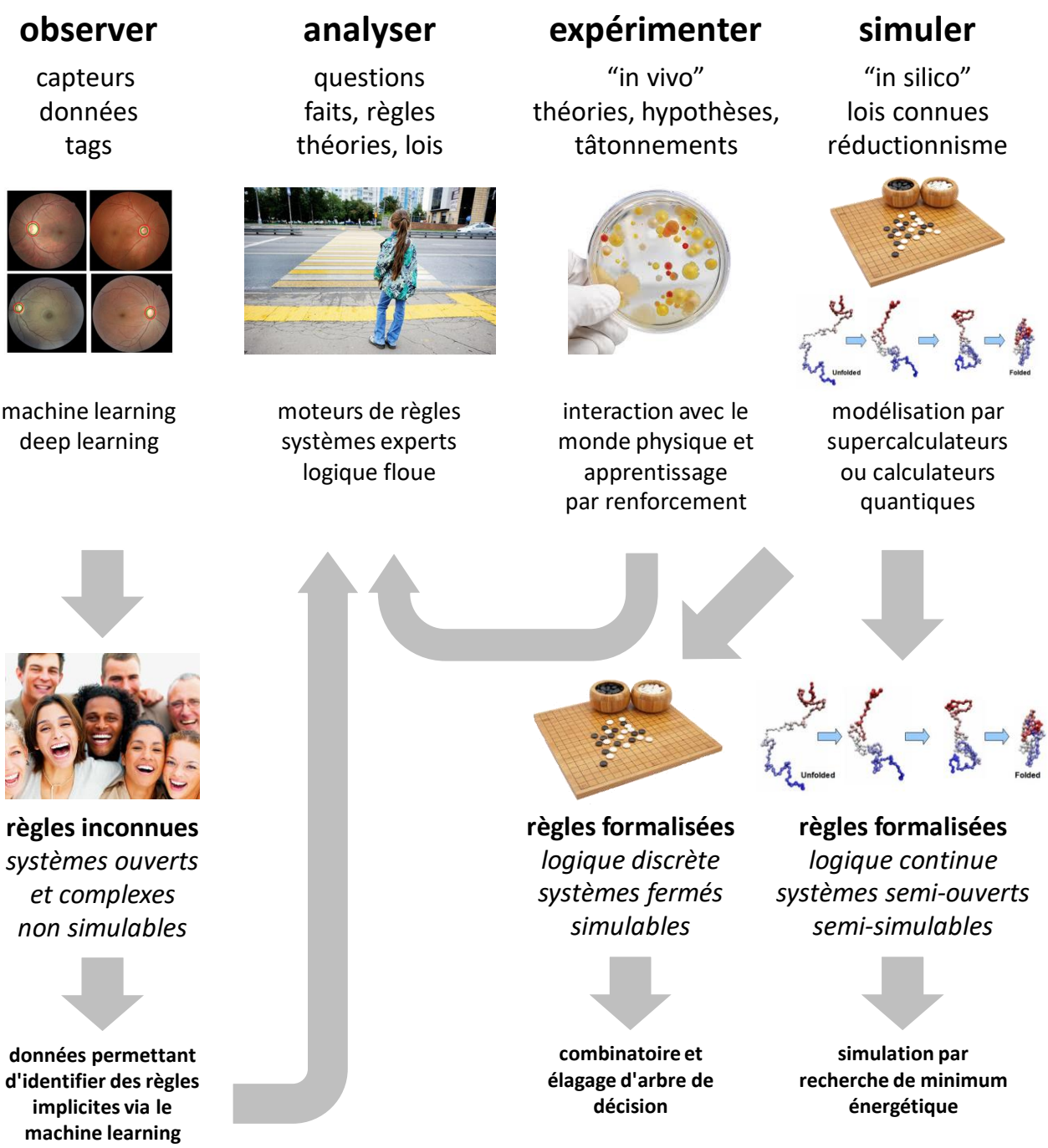
---

[38] Voir AI is about machine reasoning – or why machine learning is just an fancy plugin, de la startup allemande Arago, 2017.

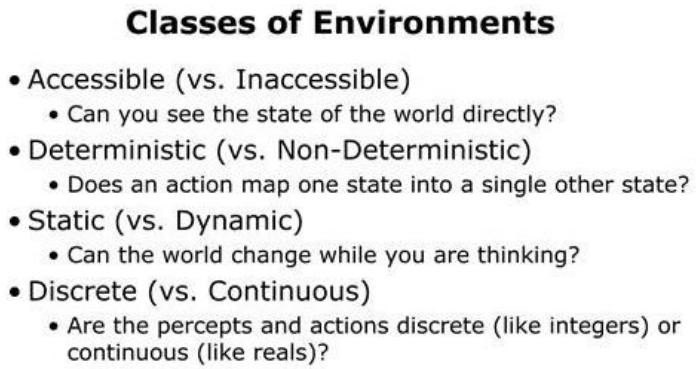
[39] Voir la partie dédiée aux théories de la complexité de ma série d'articles sur l'informatique quantique, juillet 2018.

[40] Voir à ce sujet l'intéressant Pourquoi la société ne se laisse pas mettre en équations ? de Hubert Guillaud, avril 2018.

Le scénario qui le reflète le mieux est celui de la détermination d'une politique économique d'un gouvernement. Idéalement, il faudrait mettre l'économie d'un pays et du monde en équations et simuler la modification de paramètres clés pour évaluer différentes politiques. D'un point de vue opérationnel, le nombre de paramètres est tellement important et reposant sur des dimensions sociales et psychologiques complexes, il est impossible de simuler l'ensemble. On en est réduit à opérer par tâtonnement, ou en s'accrochant aux branches d'idéologies économiques établies.



Comme le décrit la liste *ci-contre*, les paramètres d'un système à simuler ou à expérimenter sont nombreux. Est-ce que le système est accessible ou pas pour être mesurable ? Est-ce qu'il est déterministe, à savoir que des actions génèrent les mêmes conséquences à chaque itération ? Est-ce que le système est dynamique ou statique ?



Fonctionne-t-il en vase clos (comme un jeu de Go) ou dans un monde ouvert (comme pour une politique économique) ? Est-ce que les règles de fonctionnement sont discrètes (oui/non) ou continues (régies par des nombres réels) ? Tous ces paramètres vont conditionner la faisabilité et le réalisme des expérimentations et des simulations pour résoudre un problème ou prédire un état futur d'un système complexe.

Pour reboucler sur le schéma, la première colonne qui relève du machine learning illustre le fait que l'IA connexionniste est une manière de contourner l'impossibilité de simuler le monde physique. A la place, on l'observe et on utilise des méthodes probabilistes pour en déduire des règles empiriques et faire des prévisions approximatives. Mais la théorie du chaos rend le monde en grande partie imprédictible !

Etat des lieux

Faisons ici un point rapide sur l'état des lieux de l'IA en 2018. Le schéma qui suit positionne de manière rudimentaire la comparaison des capacités des machines face à celles de l'Homme. Cette approche anthropomorphique de l'IA n'est pas la seule qui soit pertinente mais elle permet de bien remettre les choses en place.

Les domaines où l'IA et les outils numériques en général dépassent déjà l'Homme sont ceux qui relèvent du calcul, comprenant les applications « data » du machine learning pour faire des prévisions ou de la segmentation, tout ce qui relève de la mémoire, notamment dans l'accès à de gros volumes d'information, les jeux de société qui sont maintenant presque tous gagnés par l'IA face à l'Homme, les derniers en date étant le Go et le Poker, et enfin, la vision artificielle, notamment celle qui est appliquée à des fonctions spécialisées comme dans le diagnostic médical.

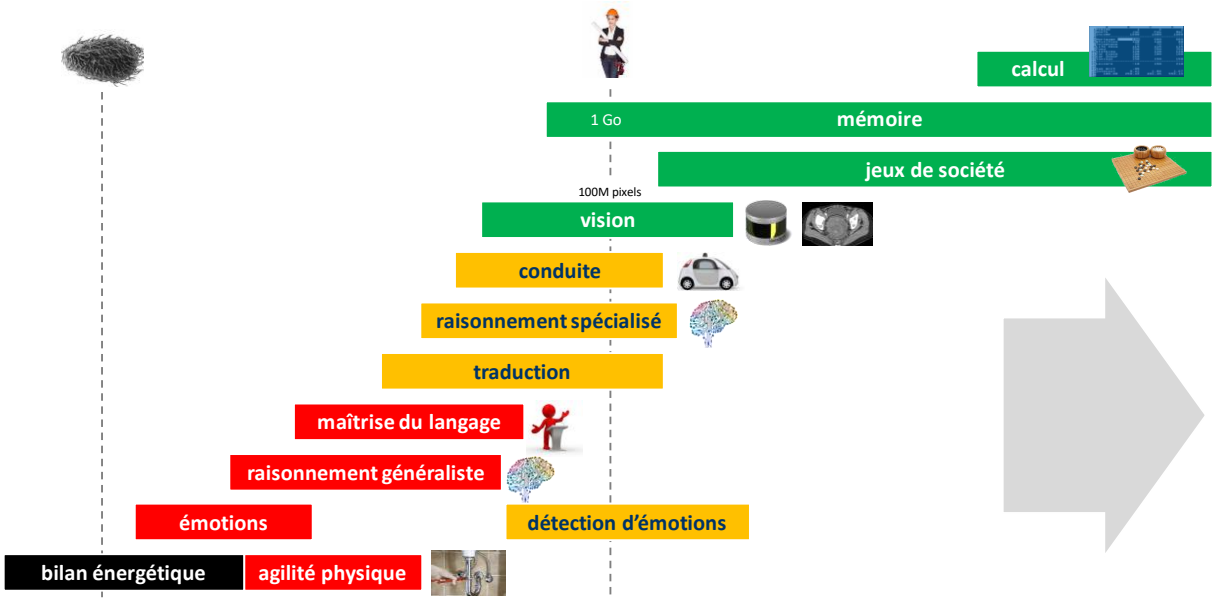
Dans les domaines où nous sommes presque ex-aequo, on peut citer la conduite autonome dans certaines conditions et le raisonnement spécialisé.

La traduction dépasse les capacités générales de l'Homme mais pas celle des spécialistes.

Pour le reste, même si les progrès sont constants, les IA ont encore du chemin à faire pour atteindre les capacités humaines. Tout ce qui relève de la maîtrise du langage est encore en deça des capacités humaines. Pour ce qui est de la traduction, un véritable bilingue fera toujours mieux qu'un système automatique.

Le raisonnement généraliste n'est pas encore possible avec l'IA actuelle[41]. Les agents vocaux rendent des services mais sont très loin de passer le test de Turing pour égaler un spécialiste humain, même Google Duplex qui prend rendez-vous avec votre docteur à votre place.

[41] Voir à ce sujet Don't believe the hype: separating AI fact from fiction de Larry Lefkowitz, juillet 2018 qui évoque la dimension mécanique de l'IA actuelle et sa difficulté à gérer du raisonnement et de la compréhension du sens.



L'agilité générale des robots reste enfin très limitée par rapport à celle de l'Homme même si l'on peut admirer les prouesses de certains robots de Boston Dynamics, qui sont d'ailleurs en partie « fake ». C'est pour cela que l'on n'est pas prêt de voir des robots s'occuper des patients dans les hôpitaux ou les maisons de retraite[42].

Deux derniers points sont encore en devenir : la gestion des émotions et le bilan énergétique. Les IA consomment en général beaucoup plus d'énergie que l'Homme avec son cerveau qui ne consomme que 20W par heure.

L'IA rend de très nombreux services au quotidien au grand public et aux entreprises, mais sans forcément relever de l'imitation de l'intelligence humaine, notamment sa capacité de raisonnement. Il en va ainsi des systèmes à base de machine learning qui essayent laborieusement de vous proposer le bon film sur Netflix ou le bon produit sur Amazon en exploitant votre historique d'usage et celui des autres Internautes.

Ces techniques de machine learning sont basées sur des méthodes et outils probabilistes qui ne correspondent pas aux sens humains. Cela permet de modérer à court terme ses frayeurs sur l'IA. Autre limitation, ils n'ont aucun moyen d'intégrer un effet de surprise dans leurs recommendations, car ils ne sont guidés que par votre consommation passée, pas par d'autres sources d'inspiration qui n'ont jamais été capturées en ligne.

Il existe d'autres segmentations de l'IA par rapport aux capacités humaines comme ces quatre catégories de systèmes : qui pensent comme des humains, qui agissent comme des humains, qui pensent rationnellement (donc, pas forcément comme des humains) et qui agissent rationnellement (itou). On peut aussi plus simplement utiliser trois catégories : les sens, le raisonnement et l'action.

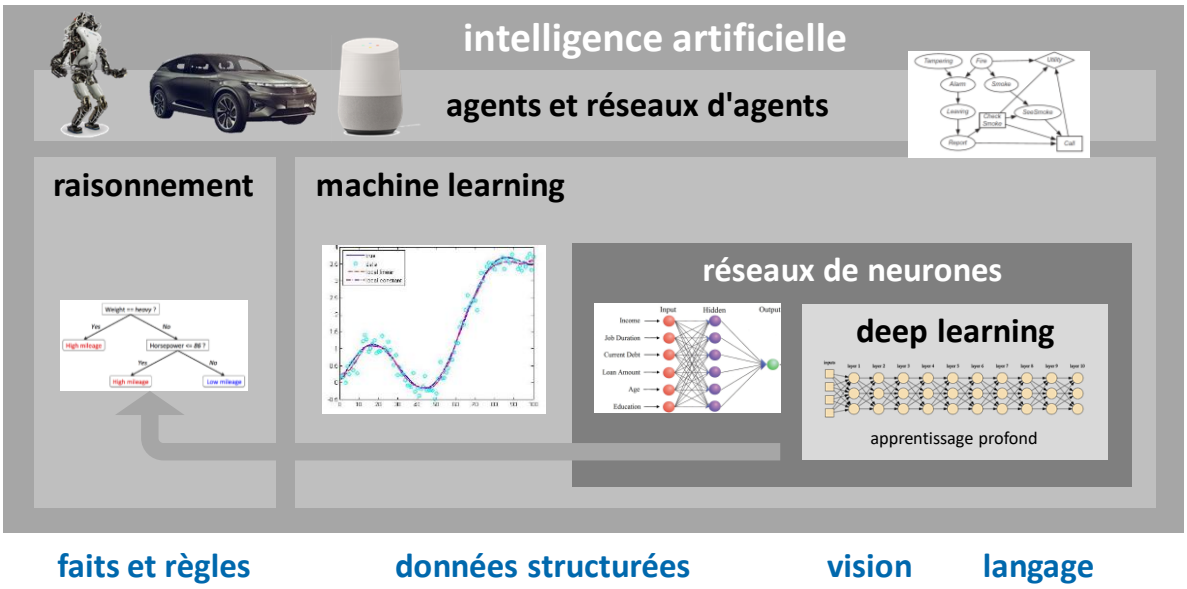
[42] Voir Progress in AI isn't as Impressive as You Might Think de Will Knight, novembre 2017 qui cite le rapport du MIT From Artificial Index Report 2017 (101 pages) piloté par Erik Brynjolfsson, le co-auteur de The Second Machine Age en 2014 et qui considère que l'on est dans une bulle de l'IA.

Algorithmes et logiciels de l'IA

Pour décrire l'univers des algorithmes, logiciels et outils de développement de l'IA, nous allons employer une segmentation simplifiée du domaine qui était déjà présente dans l'édition 2017 de ce document avec les outils du raisonnement automatique, ceux du marchine learning qui intègrent les réseaux de neurones et du deep learning et enfin, la couche d'intégration des briques de l'IA que constituent les notions d'agents et de réseaux multi-agents.

Pour chacune de ces briques fondamentales, je vais évoquer si besoin est leur ancienneté, les progrès les plus récents, les applications phares ainsi que quelques acteurs des marchés correspondants, notamment au niveau des outils de développement.

Cette segmentation couvre les principaux usages actuels de l'IA puisque l'essentiel des solutions de traitement de l'image, du langage naturel et de la perception relèvent du deep learning, celles qui gèrent des données structurées et notamment de la prévision relèvent du machine learning de base et celles qui relèvent du raisonnement et de la planification relèvent de différentes variations de moteurs de règles, solvers et outils associés.



De nombreux pans de l'IA qui ne figurent pas dans ce schéma exploitent ses différentes briques :

- Les **algorithmes évolutionnaires** ou génétiques qui peuvent s'appuyer sur du deep learning et qui testent plusieurs versions de solutions pour ne conserver que les meilleures.
- La **représentation de connaissances** qui les extrait de données textuelles non structurées, par exemple, via du deep learning, et les exploite ensuite le plus souvent dans du raisonnement automatique avec les outils de l'IA symbolique. C'est le sens de la flèche latérale dans le schéma. Mais cette représentation des connaissances, utilisée notamment dans les agents conversationnels peut être entièrement exploitée dans des réseaux de neurones, notamment des réseaux à mémoire (LSTM et variantes).
- L'**IA affective** qui exploite une panoplie large d'outils du machine learning et du deep learning voire du raisonnement automatique pour capter et classifier des éléments extérieurs des émotions humaines et agir en conséquence.
- Le **transfer learning**, ou apprentissage par transfert, est une variante du deep learning qui permet d'entraîner un réseau de neurones à partir d'un réseau de neurones déjà entraîné pour le compléter, le mettre à jour ou l'utiliser dans un domaine voisin du domaine initial.

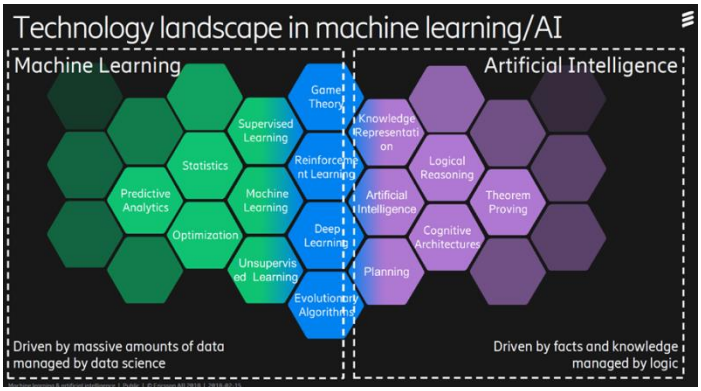
Reprenons ces grandes briques une par une :

- Les **briques du raisonnement et de la planification** qui permettent de résoudre des problèmes de logique où l'on dispose d'éléments d'informations factuels sous forme de règles, faits et contraintes, dont il faut disposer au préalable pour les exploiter. Ces problèmes peuvent relever de la maintenance, de la manière de gérer la conduite d'un véhicule autonome, du fonctionement d'un robot ou d'une machine à commande numérique dans une usine ou de l'optimisation de l'exploitation de ressources multiples pour accomplir des tâches. Les outils les plus connus dans ce domaine sont les systèmes experts bâtis avec des solveurs qui exploitent des bases de règles formelles et de faits. Les moteurs de règles s'appellent maintenant les BRMS pour Business Rules Management Systems et sont souvent intégrés dans des DMS, pour Decision Management Systems. Ces systèmes peuvent intégrer de la logique floue pour gérer des faits et règles imprécis.
- Le **machine learning** ou apprentissage automatique sert à faire des prévisions, de la classification et de la segmentation automatiques en exploitant des données en général multidimentionnelles, comme une base de données clients ou un log de serveur Internet. Le machine learning relève d'une approche probabiliste. Les outils du machine learning servent à exploiter les gros volumes de données des entreprises, autrement dit le « big data ». Le machine learning peut s'appuyer sur des réseaux de neurones simples pour les tâches complexes portant sur des données multidimentionnelles.
- Les **réseaux de neurones** constituent un sous-domaine du machine learning pour réaliser des tâches identiques, mais lorsque l'espace probabiliste géré est plus complexe. Ce biomimétisme élémentaire est exploité lorsque la dimension du problème à gérer est raisonnable. Sinon, on passe rapidement au deep learning, notamment pour le traitement de l'image et du langage.
- Le **deep learning** ou apprentissage profond, permet d'aller plus loin que le machine learning pour reconnaître des objets complexes comme les images, l'écriture manuscrite, la parole et le langage. Le deep learning exploite des réseaux de neurones multicouches, sachant qu'il en existe de très nombreuses variantes. Ce n'est cependant pas la solution à tous les problèmes que l'IA cherche à traiter[43]. Le deep learning permet aussi de générer des contenus ou d'améliorer des contenus existants, comme pour colorier automatiquement des images en noir et blanc. Deep veut dire profond. Mais le deep learning ne réfléchit pas. Il est profond parce qu'il exploite des réseaux de neurones avec de nombreuses couches de filtres. C'est tout ! Par contre, le deep learning n'est pas dédié exclusivement au traitement de l'image et du langage. Il peut servir dans d'autres environnements complexes comme en génomique. On l'utilise aussi dans des approches dites multimodales qui intègrent différents sens comme la vision et le langage.
- Les **réseaux d'agents ou systèmes multi-agents** sont un domaine méconnu qui couvre la science de l'orchestration des briques techniques de l'IA pour créer des solutions complètes. Un chatbot comme un robot est toujours un assemblage hétéroclite des briques du dessous avec des moteurs de règles, du machine learning et plusieurs techniques de deep learning. Les réseaux d'agents sont à la fois des objets conceptuels et des outils d'assemblage de briques logicielles de l'IA. Le principe d'un agent est qu'il est conceptuellement autonome, avec des entrées et sorties. L'assemblage d'agents dans des réseaux multi-agents est une version « macro » de la création de solutions d'IA.

Cette classification n'est pas la seule du marché. Il y en a quasiment autant que de spécialistes et non spécialistes du domaine de l'IA.

[43] Voir Deep learning is not AI future, de Fabio Ciucci publié en août 2017.

Cette classification issue d'une présentation d'**Ericsson** sépare le machine learning de l'intelligence artificielle qui est là, attachée au raisonnement automatique, correspondant à la première case de mon schéma[44]. C'est une vision historique, pré-2012 de l'IA, réduite à l'IA symbolique, et réductrice puisque le deep learning, qui est un sous-ensemble du marchine learning, réalise des tâches qui imitent des composantes de l'intelligence humaine comme la vision et la compréhension du langage.



Force brute et arbres de recherche

La force brute est l'inverse métaphorique de l'intelligence. C'est un moyen courant de simuler l'intelligence humaine ou de la dépasser. Pour un jeu comme les échecs, elle vise à tester toutes les possibilités et à identifier les chemins les plus optimums parmi des zillions de combinaisons.

La force brute n'est opérationnelle que si la combinatoire à tester reste dans l'enveloppe de puissance de l'ordinateur. Si elle est trop élevée, des méthodes de simplification des problèmes et de réduction de la combinatoire sont nécessaires. On utilise alors des algorithmes d'élagage qui évacuent les "branches mortes" de la combinatoire ne pouvant aboutir à aucune solution. C'est là qu'intervient une vague forme d'intelligence, mais qui repose sur une véritable force brute tout de même. Elle est systématique et pas intuitive.

C'est d'ailleurs plus facile à réaliser aux échecs qu'au jeu de Go. La combinatoire du premier est plus faible que celle du second du fait de la taille de la grille de jeu qui est respectivement de 8x8 et 19x19 cases (*ci-contre*) !

8x8

COMPLEXITY NUMBERS

Go	*40*
Chess	*14*
Scrabble	*10*
Poker	*10*
Backgammon	*8*
Checkers	*8*
Hearts	*5*
Blackjack	*2*
Craps	*0.001*
Lotteries	*0.0000001*
Roulette	*0*

19x19

La force brute a été notamment utilisée pour gagner aux échecs avec l'ordinateur **Deep Blue** d'IBM en 1997, en évaluant 200 millions de positions par seconde.

IBM Deep Blue (1996)
200 millions de positions testées par secondes
510 processeurs

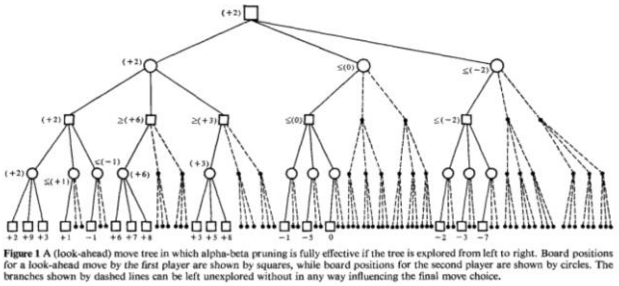


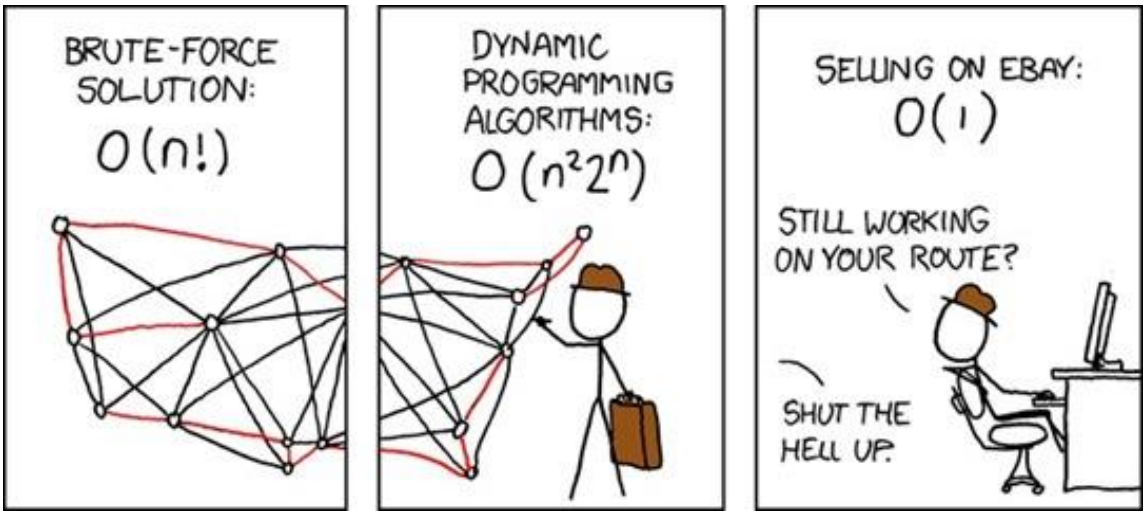
Figure 1 A (look-ahead) move tree in which alpha-beta pruning is fully effective if the tree is explored from left to right. Board positions for a look-ahead move by the first player are shown by squares, while board positions for the second player are shown by circles. The branches shown by dashed lines can be left unexplored without in any way influencing the final move choice.

Des réseaux neuronaux ont été exploités pour gagner au Go avec la solution AlphaGo de **DeepMind**, la filiale d'IA de Google. AlphaGo exploite ainsi un mélange de force brute et de deep learning permettant de faire des économies de combinatoires à tester pour identifier les meilleurs coups.

[44] Voir Ericsson Press Seminar @MWC2018, février 2018 (12 slides).

La combinatoire du jeu de Go est en effet de plusieurs ordres de grandeur supérieure à celle des échecs. AlphaGo bénéficie aussi d'un aprentissage supervisé par l'exploitation de parties de Go existantes, et d'un apprentissage par renforcement, le système apprenant en jouant contre lui-même. Dans sa version d'octobre 2017, AlphaGo Zero élague son arbre de recherche de positions avec une architecture plus simple et unifiée à base de réseaux de neurones récursifs[45].

La force brute est utilisée dans de nombreux domaines comme dans les moteurs de recherche ou la découverte de mots de passe simples. On peut considérer que de nombreux pans de l'IA l'utilisent, même lorsqu'ils s'appuient sur des techniques modernes de deep learning ou de machine learning que nous traiterons plus loin.



[46]

La force brute s'est aussi généralisée parce que la puissance des ordinateurs le permet : ils tournent plus vite, sont distribuables, le stockage coûte de moins en moins cher, les télécommunications sont abordables et les capteurs de plus en plus nombreux, des appareils photo/vidéo des smartphones aux capteurs d'objets connectés divers.

Méthodes statistiques

Les méthodes statistiques et notamment bayésiennes permettent de prévoir la probabilité d'événements en fonction de l'analyse d'événements passés.

Les réseaux bayésiens utilisent des modèles à base de graphes pour décrire des relations d'interdépendances statistiques et de causalité entre facteurs (exemple *ci-dessous*).

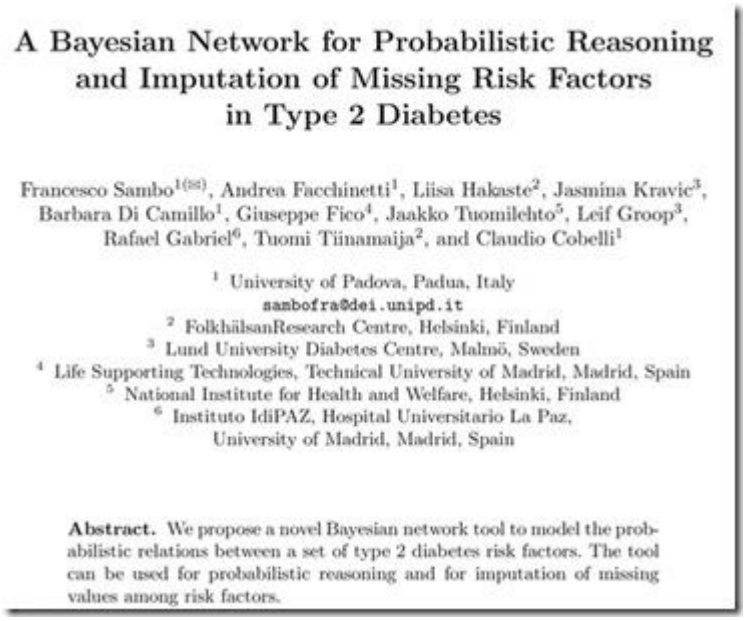
Les applications sont nombreuses comme la détection de potentiel de fraudes dans les transactions de cartes bancaires ou l'analyse de risques d'incidents pour des assurés. Elles sont aussi très utilisées dans les moteurs de recherche au détriment de méthodes plus formelles, comme le rappelle **Brian Bannon** en 2009 dans **Unreasonable Effectiveness of Data**.

La plupart des études scientifiques dans le domaine de la biologie et de la santé génèrent des corpus sous forme de résultats statistiques comme des gaussiennes d'efficacité de nouveaux médicaments. L'exploitation de la masse de ces résultats relève aussi d'approches bayésiennes.

---

[45] La méthode relève toujours de l'élagage d'arbre de décisions dans les options de jeu avec un réseau de neurones qui s'améliore par renforcement en jouant contre lui-même. Voir l'article technique Mastering the Game of Go without Human Knowledge (42 pages) qui documente la prouesse et AlphaGo Zero: Learning from scratch, de DeepMind, qui vulgarise la performance. J'ai décortiqué AlphaGo Zero dans Les conséquences pratiques d'AlphaGo Zero en novembre 2017.

[46] Source de l'image : https://xkcd.com/399/.

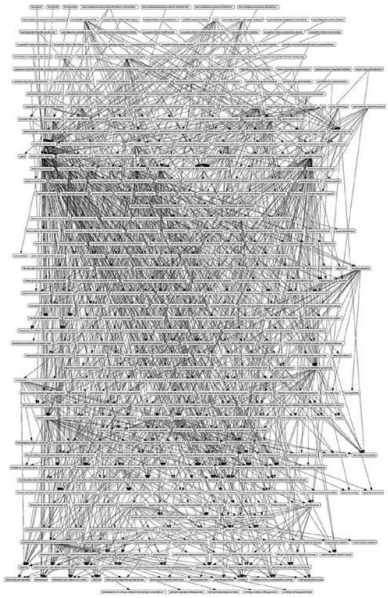


A Bayesian Network for Probabilistic Reasoning and Imputation of Missing Risk Factors in Type 2 Diabetes

Francesco Sambo[1(✉)], Andrea Facchinetti[1], Liisa Hakaste[2], Jasmina Kravic[3], Barbara Di Camillo[1], Giuseppe Fico[4], Jaakko Tuomilehto[5], Leif Groop[3], Rafael Gabriel[6], Tuomi Tiinamaija[2], and Claudio Cobelli[1]

[1] University of Padova, Padua, Italy
sambofra@dei.unipd.it
[2] FolkhälsanResearch Centre, Helsinki, Finland
[3] Lund University Diabetes Centre, Malmö, Sweden
[4] Life Supporting Technologies, Technical University of Madrid, Madrid, Spain
[5] National Institute for Health and Welfare, Helsinki, Finland
[6] Instituto IdiPAZ, Hospital Universitario La Paz, University of Madrid, Madrid, Spain

Abstract. We propose a novel Bayesian network tool to model the probabilistic relations between a set of type 2 diabetes risk factors. The tool can be used for probabilistic reasoning and for imputation of missing values among risk factors.



Le cerveau met d'ailleurs en œuvre une logique bayésienne pour ses propres prises de décision, notamment motrices, les centres associés étant situés dans le cervelet tandis que dans le cortex cérébral gère la mémoire et les actions explicites, y compris le déclenchement de mouvements[47].

Les méthodes statistiques se sont fondues avec le temps dans les techniques du machine learning et du deep learning. Ces dernières reposent en effet tous sur des modèles probabilistes pour identifier des objets ou prédire le futur[48]. S'appuyant sur de la logique formelle, l'IA symbolique ne relève pas des probabilités. Mais des combinaisons sont possibles comme lorsque l'on intègre la notion de logique floue aux moteurs de règles.

Cette partie mériterait sans doutes d'être développée !

Raisonnement automatique

Le raisonnement automatique fait partie du vaste champ de l'IA symbolique appliquant de la logique formelle. Cela faisait historiquement partie du champ de la **GOFAI**, pour "Good old fashion AI ». Elle s'opposait à l'approche connexionniste qui exploite le biomimétisme et les réseaux de neurones dans une approche probabiliste. L'approche symbolique appliquée au raisonnement automatique est plus rigoureuse mais difficile à mettre en œuvre et à généraliser car il est très difficile de collecter les règles d'un métier ou d'un domaine donné.

La formalisation du raisonnement humain remonte à **Aristote** et à l'identification de règles formelles utilisées dans l'argumentation philosophique, à base de sillogisme associant deux prémisses et une déduction (si A et B sont vrais alors C est vrai).

Suivirent les travaux de **Georges Boole** au 19e siècle et son algèbre formalisant l'usage de règles de raisonnement, puis de nombreux développements théoriques, notamment autour de la logique formelle, des calculs de prédicats, de la logique du premier et du second ordre[49].

[47] Voir Le cerveau statisticien : la révolution bayésienne en sciences cognitives, de Stanislas Dehaene, cours du Collège de France (31 slides).

[48] Voir Traditional statistical methods often out-perform machine learning methods for time-series forecasts de Paul Cuckoo, juillet 2018.

[49] Voir Intelligence Artificielle Symbolique de Guillaume Piolle, 2015.

A vrai dire, on ne sait toujours pas dire si le raisonnement humain s'appuie sur une manipulation de symboles à haut niveau ou par assemblage de connexions de bas niveau dans le cerveau, que l'on appelle approche connexioniste ou sous-symbolique.

Les outils du raisonnement automatique en IA sont très divers avec les moteurs de règles, les arbres d'exploration et autres graphes, les ontologies et les outils de modélisation des connaissances. Pendant longtemps, les approches symboliques et connexionnistes se sont opposées. Les recherches les plus récentes en IA visent à les rapprocher. D'au moins deux manières. Tout d'abord en exploitant le deep learning pour extraire des règles de documents en texte libre ou en associant textes et images dans des logiques multimodales, qui alimentent ensuite des moteurs de règles traditionnels. Et puis en cherchant à intégrer des briques de raisonnement dans le deep learning, le raisonnement étant l'aboutissement de techniques (toujours probabilistes) de traitement du langage et de gestion des connaissances.

Démonstrations de théorèmes

Les débuts des moteurs de règles à la base des systèmes experts remontent à 1957 quand **Alan Newell** et **Herbert Simon** développaient le General Problem Solver (GPS), un logiciel de résolution de problèmes mathématiques utilisant des règles modélisant les inférences possibles d'un domaine et résolvant un problème en partant de la solution attendue et en remontant vers les hypothèses. Ce GPS faisait suite au **Logic Theorist** produit début 1956 par ces mêmes Allen Newell, Herbert Simon et Cliff Shaw, quelques mois avant la création d'appellation d'intelligence artificielle par Jim MacCarthy. Le Logic Theorist démontrait quelques théorèmes mathématiques à base de logique axiomatique[50].

En 1958, **John McCarthy** créait le langage de programmation LISP adapté en particulier à la création de moteurs de règles. Il a abouti à la création d'une petite industrie dans les années 1980 avec les ordinateurs spécialisés de **Lisp Machines** et **Symbolics** (1979-2005) et les logiciels d'**Intellicorp**[51].

En 1959, **Herbert Gelernter** créait le Geometry Theorem Prover[52], capable de démontrer des théorèmes de géométrie et tournant sur un IBM 704. Le même Gelernter est à l'origine du SYNCHEM (SYNthetic CHEMistry), un système expert des années 1970 capable de déterminer des réactions chimiques de synthèse de molécules organiques.

Dans les années 1970, **Robert Kowalski** de l'Université d'Edimbourg fit avancer les travaux dans la programmation logique. Puis les français **Alain Colmerauer** et **Philippe Roussel** créèrent le langage de programmation **Prolog** (**Pro**grammation en **log**ique) qui connut ses heures de gloire pendant les années 1980 et 1990, et surtout dans l'éducation et la recherche. Prolog était principalement utilisé en Europe tandis que LISP l'était de manière dominante aux USA. Un *replay* de la bataille NTSC vs PAL/Secam dans la TV couleur !

```
mother_child(trude, sally).

father_child(tom, sally).
father_child(tom, erica).
father_child(mike, tom).

sibling(X, Y)      :- parent_child(Z, X), parent_child(Z, Y).

parent_child(X, Y) :- father_child(X, Y).
parent_child(X, Y) :- mother_child(X, Y).
```

This results in the following query being evaluated as true:

```
?- sibling(sally, erica).
Yes
```

[50] Voir le compte-rendu des auteurs sur Logic Theorist : The Logic Theory Machine A Complex Information Processing System, juin 1956 (40 pages).

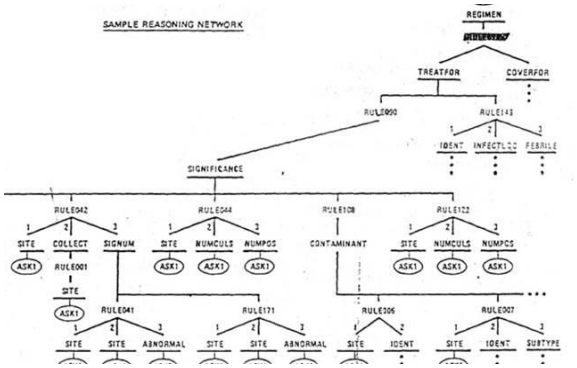
[51] Créé en 1980 et maintenant spécialisé dans les logiciels de gestion d'applications pour SAP, un métier plus terre à terre

[52] Voir Realization of a Geometry-Theorem Proving Machine (30 pages).

Prolog est un langage déclaratif qui sert à gérer de la logique du premier ordre avec la déclaration de relations, faits et règles. Il est notamment utilisé pour l'analyse du langage. Prolog est d'ailleurs utilisé à cet effet dans DeepQA, une des briques d'IBM Watson.

Premiers systèmes experts

Quelques expérimentations ont marqué les débuts des systèmes experts comme **MYCIN**, un système permettant de déterminer les bactéries responsables d'infections en fonction des symptômes (*ci-dessous*) avec une base de 450 règles.




```
RULE036
PREMISE: ($AND (SAME CNTXT GRAM GRAMNEG)
               (SAME CNTXTM MORPH ROD)
               (SAME CNTXT AIR ANAEROBIC))
ACTION: (CONCLUDE CNTXT IDENTITY BACTEROIDES TALLY 0.6)

IF: 1) The gram stain of the organism is gramneg, and
    2) The morphology of the organism is rod, and
    3) The aerobicity of the organism is anaerobic

THEN: There is suggestive evidence (0.6) that the identity
           of the organism is bacteroides
```

Les systèmes experts ont été théorisés dans le cadre du **Stanford Heuristic Programming Project** en 1980. Ils répondent à des questions dans des domaines spécifiques dont on a codifié la connaissance. Cela permit à l'IA de se rendre utile dans des domaines spécifiques, comme dans la santé ou la maintenance dans l'industrie[53].

Les moteurs de règles et les solveurs sont couramment employés dans les systèmes experts depuis les années 1980[54]. Et ils ont connu de nombreux progrès (*ci-dessous*) malgré l'hiver de l'IA de la fin des années 1980 et débuts 1990. C'était surtout un hiver des systèmes experts et du LISP !

DENDRAL – composition de matériaux, Feigenbaum, Buchanan et al, 1965
PROSPECTOR – prospection géologique, 1977
R1 – Digital Equipment, 1982
DIPMETER – prospection géologique, 1982
MYCIN – diagnostic de maladies infectieuses, 1983
SOAR - Laird, 1983
Cyc – SI généraliste, Lenat and Guha, 1984

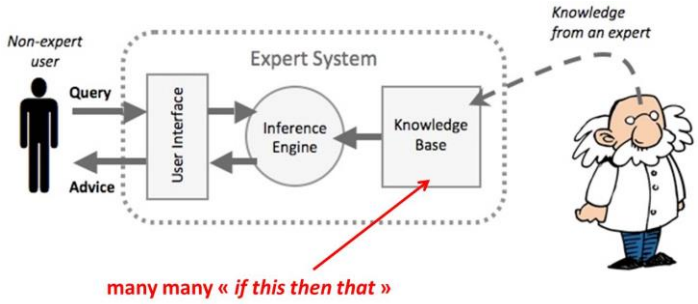
ILOG Rules - devenu IBM Decision Optimization Manager, 1997
EPIC - trafic aérien, Rosbe, Chong, and Kieras, 2001
Web sémantique / RDF - 2001
ACT-R - Anderson and Lebiere, 2003
ICARUS - Langley, 2005
SNePS - Semantic Network Processing System, Shapiro, 2007

Les moteurs de règles s'appuient sur la notion de raisonnement contraint par des règles et exploitant des bases de faits. On fournit au moteur un ensemble de règles et de faits pouvant par exemple représenter le savoir des experts dans un domaine donné. Avec des règles proches de la programmation logique du genre "*si X et Y sont vrais, alors Z est vrai*" ou "*X entraîne Y*".

On peut alors interroger le système en lui posant des questions genre "*est-ce que W est vrai ?*" et il va se débrouiller pour exploiter les règles enregistrées pour répondre à la question. Les moteurs de règles utilisent la théorie des graphes et la gestion de contraintes.

[53] Comme le système expert français SACHEM (Système d'Aide à la Conduite des Hauts fourneaux En Marche) chez Arcelor-Mittal qui pilote les hauts-fourneaux en exploitant les données issues de capteurs. Le projet a été mené entre 1991 et 1998 et serait toujours en opération.

[54] On peut citer notamment l'outil de développement Nexpert pour Macintosh et PC de **Neuron Data**, une startup créée en 1985 aux USA par les français Alain Rappaport, Patrick Perez et Jean-Marie Chauvet. Elle a été revendue au début des années 2000à l'Allemand Brokat puis à différents acquéreurs successifs avant de disparaître.



Un système expert s'appuie sur deux composantes clés : une **base de connaissance**, générée souvent manuellement ou éventuellement par exploitation de bases de connaissances existantes, et un **moteur de règles**, plus ou moins générique, qui va utiliser la base de connaissance pour répondre à des questions précises.

Les systèmes experts peuvent expliquer le rationnel de leur réponse. La traçabilité est possible jusqu'au savoir codifié dans la base de connaissances, un avantage que les réseaux de neurones du deep learning et le machine learning n'ont pas encore.

Systèmes experts d'aujourd'hui

On compte encore des outils et langages dans ce domaine et notamment l'offre du français **ILOG**, acquis en 2009 par IBM et dont les laboratoires de R&D sont toujours à Gentilly près de Paris, au sud du boulevard Périphérique. Le moteur d'inférence ILOG JRules est devenu **IBM Operational Decision Manager**. De son côté, ILOG Solver est une bibliothèque C++ de programmation par contraintes, devenue IBM ILOG CPLEX CP Optimizer. Une stratégie de branding moins efficace que celle d'IBM Watson, comme nous le verrons bien plus loin.

La mise en place de systèmes experts se heurtait à la difficulté de capter la connaissance des experts. Les temps de calcul pour les faire fonctionner étaient également longs avec les ordinateurs de l'époque. La loi de Moore a permis de limiter ce dernier écueil depuis.

Il existe d'autres types de systèmes experts qui mettent en œuvre la notion de programmation par contraintes, permettant d'atteindre un objectif en fonction d'une base de règles, d'objectifs et de contraintes opérationnelles. Dans de nombreux domaines, la force brute et le deep learning se sont ensuite imposés en lieu et place de la logique formelle et de la captation manuelle de connaissances.

moteurs de règles

open source

CLIPS : moteur de règles dans le domaine public.

Drools : distribué par Red Hat.

DTRules : moteur de règles en Java.

Gandalf : moteur de règles tournant sur PHP.

OpenL Tablets : business centric rules and BRMS.

propriétaires

Corticon : moteur de règles sous Java et .NET, filiale de Progress Software.

IBM Operational Decision Manager : ex ILOG Rules.

JESS : moteur de règle Java, sur-ensemble du langage CLIPS.

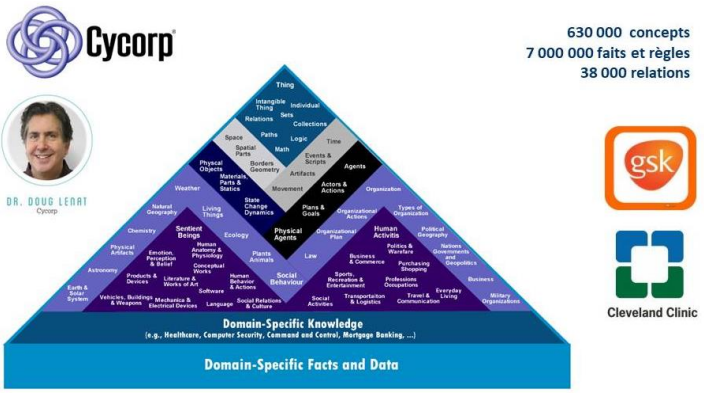
Microsoft Azure Business Rules Engine : framework de moteur de règle en .NET.

Oracle Policy Automation : modélisation et déploiement de règles.

Les logiciels de moteurs de règles du marché sont appelé BRMS pour **Business Rules Management Systems**. L'offre est assez abondante mais plus ancienne et moins connue que celle qui concerne le machine learning et le deep learning (*ci-dessous*). Cette offre de BRMS est maintenant intégrée dans le concept plus large de **Decision Management Systems** qui associent des moteurs de règles et des outils d'*analytics*.

L'un des systèmes experts les plus ambitieux des années 1980 était **Cyc**. Il devait comprendre une énorme base de connaissances de centaines de milliers de règles. Ce projet était piloté par Doug Lenat du consorsium de recherche privé MCC qui ferma ses portes en 2000.

Doug Lenat l'a transformé en projet entrepreneurial avec **Cycorp**, lancée en 1994[55]. Cette dernière propose une base de connaissance intégrant 630 000 concepts, 7 millions de faits et règles et 38 000 relations, le tout étant exploitable par des moteurs de règles (*ci-contre*). La base est notamment alimentée par l'analyse de documents disponibles sur Internet. Mais ce projet est considéré comme un échec.

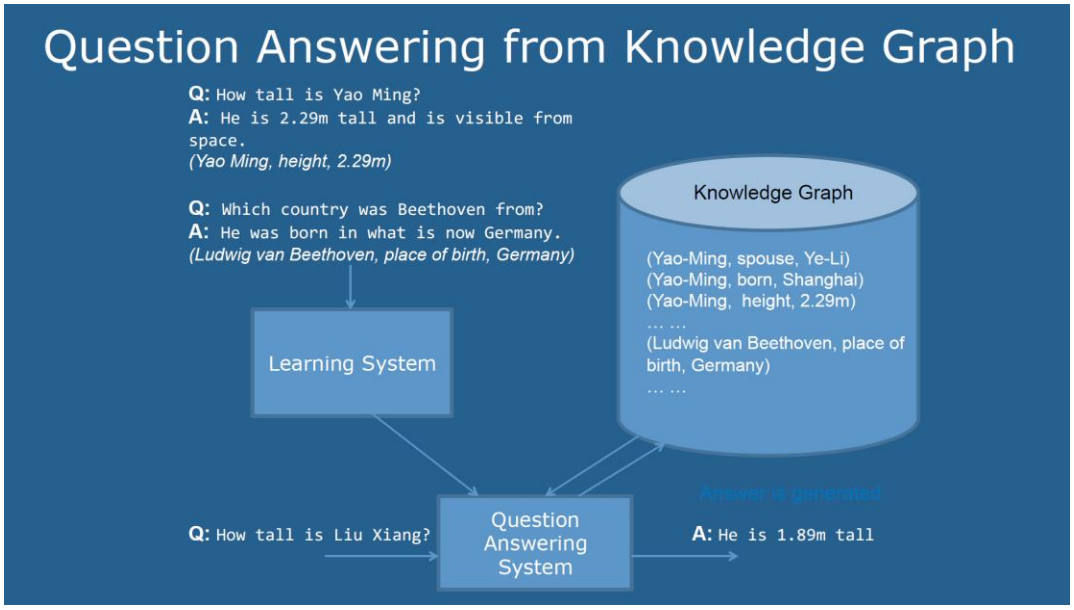


Cycorp est une sorte de laboratoire de recherche privé en IA financé par des contrats du gouvernement US, dont la DARPA, et d'entreprises privées. Il propose une suite d'outils en open source et licence commerciale permettant d'exploiter des dictionnaires, ontologies et bases de connaissances pour répondre à des questions d'analystes.

Le système expert OpenCyc 4.0 qui exploitait la base de Cycorp n'est plus disponible en open source depuis 2017 pour éviter le fameux phénomène de la fragmentation du code (« fork »). Il est depuis commercialisé sous forme de licences dédiées à la recherche ou de licences commerciales. Cycorp est devenu une société de conseil et d'intégration en IA, spécialisée notamment dans le traitement du langage et des connaissances.

L'initiative open source **Schema.org** lancée par Google, Microsoft, Yahoo et le Russe Yandex propose de son côté des millions de types, descriptions de faits exploitables par les moteurs de recherche et les moteurs de règles. Il permet notamment aux web masters de standardiser la nomenclature utilisée dans les descriptifs de pages web.

Les outils dotés de capacités de raisonnement continuent d'évoluer pour faire avancer le champ de la représentation des connaissances et du raisonnement. Les techniques associées sont moins connues que celles du machine learning et du deep learning, ce d'autant plus qu'elles sont de plus en plus hybrides. Ainsi, un moteur de règles peut-il exploiter des règles elles-mêmes générées par analyse du langage dans des réseaux de neurones récurrents.

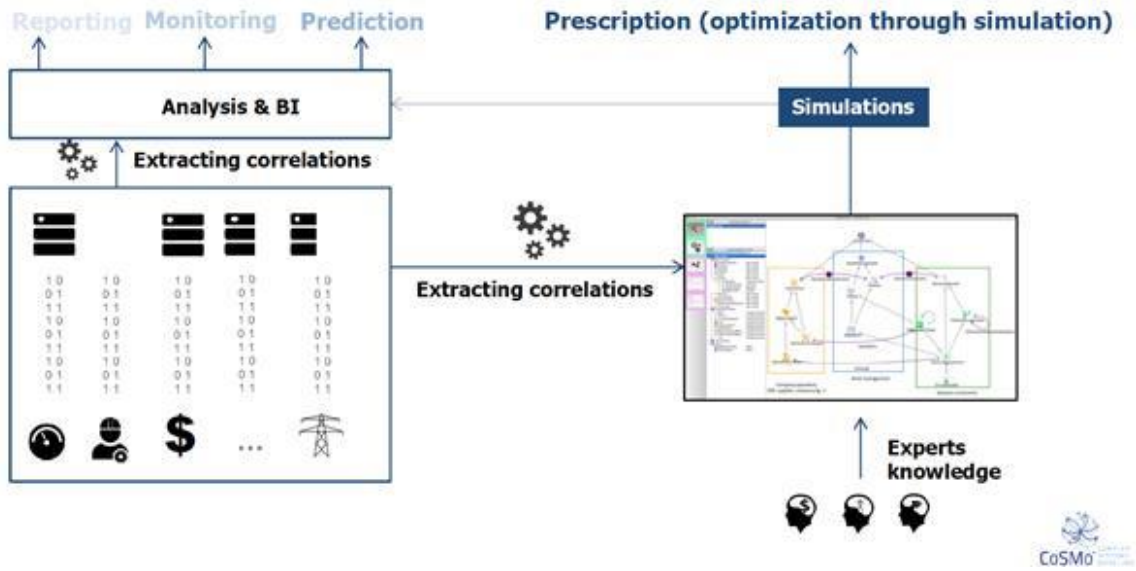


[55] Voir ce talk à TEDx Youth Austin de Doug Lenat qui date de 2015 (16 mn).

Le deep learning et les réseaux de neurones récurrents que nous verrons plus loin alimentent maintenant couramment les bases de connaissances et les moteurs de règles qu'ils ont contribué indirectement à faire décliner ! Les bases de connaissances sont devenues des **Knowledge Graphs**, mais c'est la même chose, comme l'illustre le schéma *ci-dessus* issu de Toward Neural Symbolic Processing de Hang Li, 2017 (36 slides).

Les systèmes experts peuvent aussi être alimentés par l'exploitation de données opérationnelles (big data, machine learning). C'est l'approche de **Cosmo Tech** (2010, France/USA, $30,3M), une startup spin-off de l'ENS Lyon et du CNRS basée à Lyon et aux USA qui a développé une plateforme logicielle de modélisation et de simulation de systèmes complexes. Elle s'appuie sur le langage de modélisation CosML qui sert à représenter les états ainsi que les comportements des systèmes complexes et à les étudier grâce à de la simulation. Le système exploite des règles métier et des corrélations extraites de données de production via des techniques de machine learning (*schéma ci-dessous*).

La solution est déclinée dans diverses industries comme avec leur application *Asset Investment Optimization* (AIO) dédiée aux énergéticiens, *Crisis Management* qui permet la gestion de crise et *Smart Territories* qui permet de modéliser des systèmes complexes pour la ville intelligente. C'est un excellent exemple d'hybridation technologique illustrant la manière dont les systèmes experts s'intègrent dans les solutions d'IA.



On peut aussi citer **Khresterion** (2014, France) qui propose un logiciel d'aide au diagnostic dans divers domaines. Ils s'appuient sur un moteur de règle, K Engine, qui exploite une représentation de la connaissance sous forme d'ontologies et adopte une structure en graphe et non d'arborescence, ce qui la rend très ouverte. Cela permet de gérer la contradiction et la non complétude d'informations. La société travaille dans les domaines financiers et juridiques après avoir tâté du domaine de la santé.

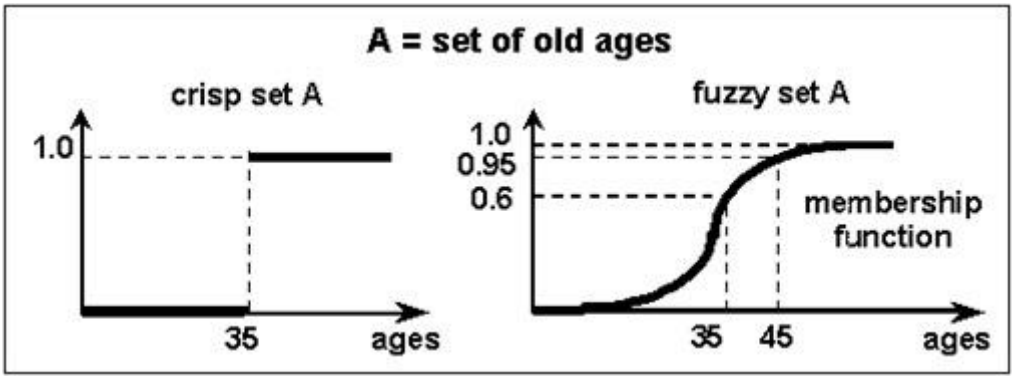
Il y aussi **ExpertSystem** (1989, Italie) qui propose diverses solutions pour les entreprises à base de systèmes experts et d'outils de traitement du langage naturel.

Mais comme nous le verrons dans la partie dédiée à l'AGI, des tentatives nombreuses visent à utiliser des réseaux de neurones pour faire du raisonnement symbolique. L'un des premiers du genre fut le **KBANN** (Knowledge-based Artificial Neural Network) de Towell et Shavlik en 1994[56] mais qui ne semble pas avoir abouti à des applications pratiques.

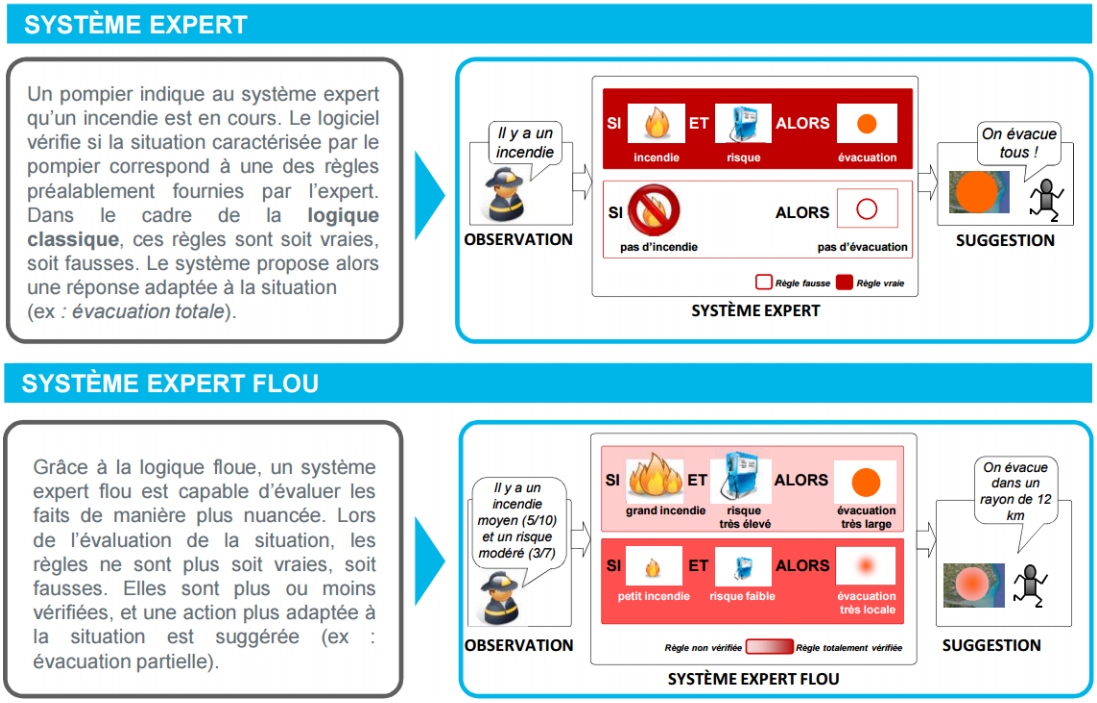
[56] Voir Knowledge-Based Artificial Neural Networks, 1994 (45 pages) qui est cité dans Reasoning with Deep Learning: an Open Challenge de Marco Lippi, 2016 (22 slides).

Logique floue

La logique floue est un concept de logique inventé par l'américain **Lofti Zadeh** ("Fuzzy Logic") en 1965[57]. Elle permet de manipuler des informations floues qui ne sont ni vraies ni fausses, en complément de la logique booléenne, mais à pouvoir faire des opérations dessus comme l'inversion, le minimum ou le maximum de deux valeurs. On peut aussi faire des OU et des ET sur des valeurs "floues".



Quid des applications de la logique floue[58]? On les trouve dans le contrôle industriel, dans des boîtes de vitesse chez **Volkswagen** (pour tenir compte de l'intention "floue" du conducteur), pour gérer des **feux de circulation** et maximiser le débit, dans de nombreuses applications du BTP, dans la reconnaissance de la parole et d'images, le plus souvent, en complément du bayésien. Des **dizaines de milliers de brevets** auraient été déposés pour protéger des procédés techniques utilisant la théorie de la logique floue.



Lotfi Zadeh
1921-2017

[57] J'avais eu l'occasion de l'entendre la présenter lors d'une conférence à l'Ecole Centrale en 1984, lorsque j'étais en option informatique en troisième année. Ça ne nous rajeunit pas ! Lotfi Zadeh est décédé en septembre 2017. Voir Lotfi Zadeh 1921–2017, par Richard Lipton et Ken Regan, octobre 2017.

[58] Quelques exemples d'applications de la logique floue dans cette présentation : Applications of fuzzy logic de Viraj Patel, 2016 (22 slides).

Les moteurs de règles de systèmes experts peuvent d'ailleurs intégrer les principes de la logique floue (*ci-dessous*).

Ceci dit, la logique floue n'est pas utilisée de manière très courante, notamment du fait que les systèmes experts ne sont plus à la mode depuis une quinzaine d'année. J'ai juste identifié la startup **Zsolusionz** qui proposait des solutions à base de logique floue mais qui n'a plus de traces sur Internet.

Planification opérationnelle

La planification est une autre branche de l'IA symbolique. Elle vise à résoudre des problèmes de logique et d'optimisation qui sont très nombreux dans les entreprises, notamment dans les utilities et les transports.

Comment allouer ses ressources pour qu'elles soient les mieux utilisées ? Comment maximiser son revenu avec ses ressources existantes ? Comment optimiser le parcours d'un livreur ou d'un commercial ?

Les solutions pour résoudre ces problèmes sont variées. Elles peuvent faire appel à des algorithmes ad-hoc, à de la simulation, à des réseaux d'agents intelligents, une méthode qui est à la frontière entre l'intelligence artificielle et l'étude des systèmes complexes[59].

Là encore, on se retrouve rapidement dans des situations d'hybridation d'algorithmes et technologies pour résoudre des problèmes qui intègrent des règles de logique et des contraintes numériques (*aka* : programmation par contrainte, logique propositionnelle, logiques monotone et non-monotone, des notions de combinatoires et d'exploration d'arbres de décision), le traitement du langage et la modélisation des connaissances.

Machine learning

Le vaste domaine du machine learning, ou apprentissage automatique, est défini comme le champ de l'IA qui utilise des méthodes probabilistes pour apprendre à partir des données sans être programmé explicitement.

D'un point de vue pratique, le machine learning vise notamment à identifier des tendances, à faire des prévisions sur des données (régressions linéaires et non linéaires), à découvrir des corrélations entre données et événements (comme pour déterminer si un logiciel est un virus, si un client risque de quitter un service sur abonnement ou au contraire, s'il sera intéressé par telle ou telle offre ou qu'un tableau clinique d'un patient est symptomatique de l'émergence d'une pathologie de longue durée), à segmenter des jeux de données (comme une base clients), à reconnaître des objets (des lettres, des objets dans des images), le tout en exploitant des données d'entraînement.

Selon son père fondateur **Arthur Samuel,** qui en définit le terme en 1959, le machine learning donne aux machines la capacité d'apprendre sans être explicitement programmées. Le machine learning requiert presque toujours de faire des choix de méthodes et des arbitrages manuels pour les data scientists et les développeurs de solutions. Le choix des méthodes reste pour l'instant manuel, même si certaines startups essayent d'automatiser ce processus.

L'apprentissage automatique s'appuie sur des données existantes. Elles lui permettent de produire des prévisions, des segmentations ou des labels à partir de la généralisation d'observations. La qualité et la distribution statistique des données d'entraînement doit permettre de faire des prévisions de bon niveau. Si les données ne représentent pas correctement l'espace du possible, les prévisions ou classifications ne seront pas bonnes et elles seront biaisées. Les données sont donc absolument critiques pour la qualité des résultats.

---

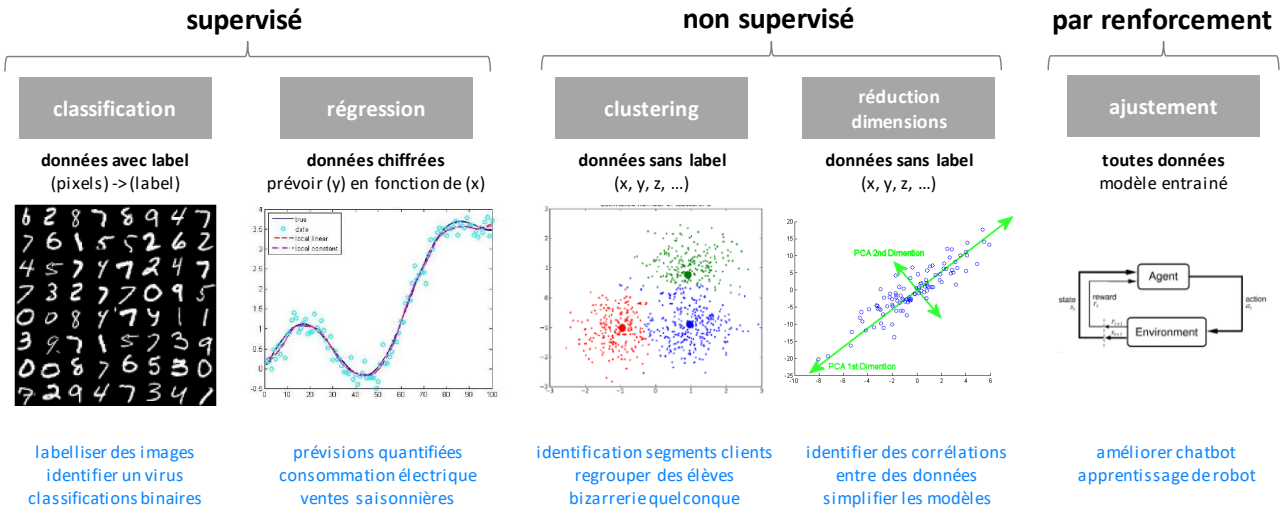
[59] Source : Renouveau de l'intelligence artificielle et de l'apprentissage automatique, un rapport de l'Académie des Technologies coordonné par Yves Caseaux et publié en avril 2018.

Un bon système de machine learning doit pouvoir s'adapter à différentes contraintes comme une évolution permanente des données d'entraînement, ainsi que leur incomplétude et leur imperfection.

On distingue en général trois grandes catégories de machine learning, reprenant le schéma *ci-dessus* :

- L'**apprentissage supervisé** avec la classification qui permet de labelliser des objets comme des images et la régression qui permet de réaliser des prévisions sur des valeurs numériques. L'apprentissage est supervisé car il exploite des bases de données d'entraînement qui contiennent des labels ou des données contenant les réponses aux questions que l'on se pose. En gros, le système exploite des exemples et acquiert la capacité à les généraliser ensuite sur de nouvelles données de production.
- L'**apprentissage non supervisé** avec le clustering et la réduction de dimensions. Il exploite des bases de données non labellisées. Ce n'est pas un équivalent fonctionnel de l'apprentissage supervisé qui serait automatique. Ses fonctions sont différentes. Le clustering permet d'isoler des segments de données spatialement séparés entre eux, mais sans que le système donne un nom ou une explication de ces clusters. La réduction de dimensions (ou embedding) vise à réduire la dimension de l'espace des données, en choisissant les dimensions les plus pertinentes. Du fait de l'arrivée des big data, la dimension des données a explosé et les recherches sur les techniques d'embedding sont très actives.
- L'**apprentissage par renforcement** pour l'ajustement de modèles déjà entraînés en fonction des réactions de l'environnement. C'est une forme d'apprentissage supervisé incrémental qui utilise des données arrivant au fil de l'eau pour modifier le comportement du système. C'est utilisé par exemple en robotique, dans les jeux ou dans les chatbots capables de s'améliorer en fonction des réactions des utilisateurs. Et le plus souvent, avec le sous-ensemble du machine learning qu'est le deep learning. L'une des variantes de l'apprentissage par renforcement est l'apprentissage supervisé autonome notamment utilisé en robotique où l'IA entraîne son modèle en déclenchant d'elle-même un jeu d'actions pour vérifier ensuite leur résultat et ajuster son comportement.

Voici un schéma maison qui résume tout cela de manière visuelle :



Nous allons explorer une par une ces différentes méthodes de machine learning et les modèles mathématiques sur lesquelles elles s'appuient.

Classification

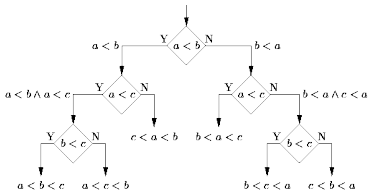
Il s'agit de pouvoir associer une donnée complexe comme une image ou un profil d'utilisateur à une classe d'objet, les différentes classes possibles étant fournies a priori par le concepteur.

La classification utilise un jeu de données d'entraînement associé à des descriptifs (les classes) pour la détermination d'un modèle. Cela génère un modèle qui permet de prédire la classe d'une nouvelle donnée fournie en entrée.

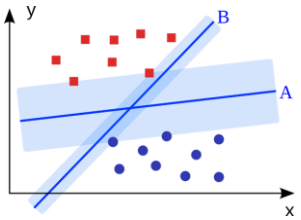
Dans les exemples classiques, nous avons la reconnaissance d'un simple chiffre dans une image, l'appartenance d'un client à un segment de clients ou pouvant faire partie d'une typologie particulière de clients (mécontents, pouvant se désabonner à un service, etc) ou la détection d'un virus en fonction du comportement ou de caractéristiques d'un logiciel.

Il existe plusieurs méthodes de classification dont voici les principales.

Les **arbres de décision** que l'on appelle aussi les CART (Classification And Regression Tree) exploitent des critères discriminants, comme dans un moteur de règles. Ils permettent de classifier un objet en se posant successivement des questions (comparaison de données, …). Il en existe plusieurs sortes, telles que les CHAID (pour CHI-square Adjusted Interaction Detection) qui peuvent utiliser des branches multiples à chaque nœud.



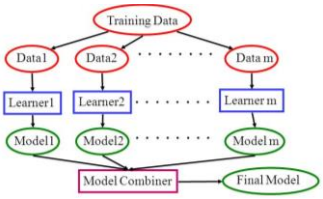
arbres de décision



Support Vector Machines

$$P(c_j|d_i)=\frac{P(c_j)\prod_{k=1}^{|d_i|}P(w_{d_{ik}}|c_j)}{\sum_{r=1}^{|C|}P(c_r)\prod_{k=1}^{|d_i|}P(w_{d_{ik}}|c_r)}$$

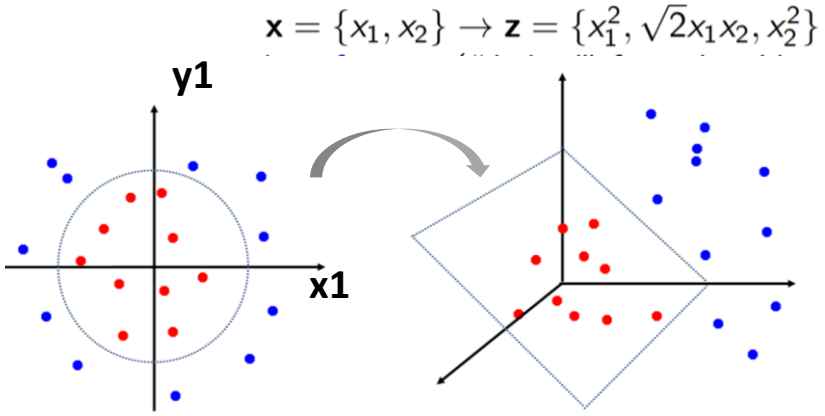
classification Bayésienne naïve



Ensemble Methods

Les **Support Vector Machines** (SVM) linéaires cherchent à identifier une droite ou un hyperplan dans le cas d'un modèle à plusieurs dimensions qui permette de distinguer les classes d'objets les unes des autres de manière binaire en essayant de les séparer par une marge, représentée par une bande autour de l'hyperplan, aussi large que possible. On les appelle aussi des modèles d'indépendance conditionnelle simples.

Les SVM peuvent utiliser un modèle **non linéaire** lorsque les objets à séparer dans l'espace ne peuvent pas être isolés de part et d'autre d'un hyperplan. On recherche alors une fonction qui va transformer ces données, par exemple en 2D (x1, x2), dans un espace à deux ou trois dimensions, les dimen¬sions étant un polynôme de x1 et x2, qui va permettre une séparation des données par un hyperplan si on ajoute une dimension[60].



Les **classifications naïves bayésiennes** utilisent les probabilités pour départager les objets dans des classes en fonction de caractéristiques bien établies[61].

---

[60] Source : Kernel Methods and Nonlinear Classification de Piyush Rai, un cours de Stanford, 2011 (95 slides) ainsi que SVM and Kernel machine - linear and non-linear classification de Stéphane Canu, 2014 (78 slides). L'exemple donné en schéma pourrait être traité avec une méthode plus simple consistant à transformer les coordonnées (x, y) en coordonnées polaires avec longueur vecteur + angle. On conserverait deux dimensions et la séparation liénaire SVM pourrait alors fonctionner.

[61] La fiche Wikipedia de la classification naïve bayésienne explique bien son principe. Elle est inspirée des travaux de Thomas Bayes (1702 – 1761) repris ensuite par Laplace. Voir également la présentation Naïve Bayes Classifier (37 slides).

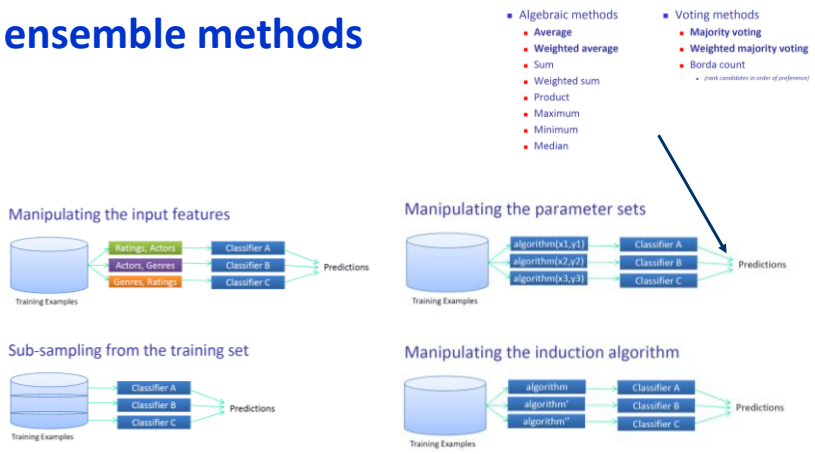
À chaque hypothèse de départ, on associe une probabilité. L'observation d'une ou de plusieurs instances peut modifier cette probabilité. On peut parler de l'hypothèse la plus probable au vu des instances observées. Les probabilités bayésiennes présupposent l'indépendance des attributs utilisés.

On retrouve cela dans le théorème de Bayes selon lequel Probabilité (A sous condition de B) = Probabilité (B sous condition de A)* Probabilité (A) / Probabilité (B). Ca fonctionne bien seulement si les probabilités de A et B sont bien indépendantes, ce qui n'est pas toujours le cas dans la vraie vie.

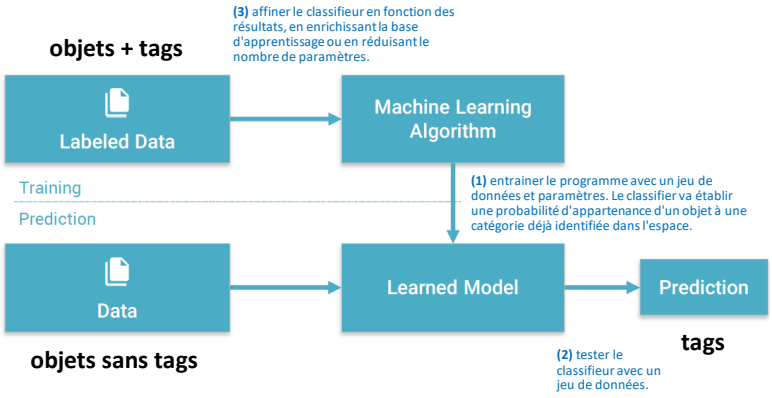
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Enfin, les **méthodes des ensembles** combinent plusieurs méthodes de classification pour en panacher les résultats et renforcer le poids des meilleures méthodes sans dépendre d'une seule d'entre elles.

Les méthodes des ensembles peuvent combiner des méthodes dites algébriques (avec une moyenne, une moyenne pondérée, un maximum, un minimum ou une médiane) et des méthodes par vote (utilisant la majorité, un vote pondéré, …), comme représenté dans le schéma *ci-contre*[62]. Le choix des assemblages dépend de la distribution statistique des données d'entraînement.



Un modèle mathématique de machine learning est entraîné avec un jeu de données d'apprentissage. Cet entraînement consiste à déterminer la bonne méthode à utiliser ainsi que les paramètres mathématiques du modèle retenu. Il va générer un modèle entraîné, et ses variables de fonctionnement pour faire une prévision ou une classification.



Le modèle entraîné est ensuite alimenté avec de nouveaux objets pour prédire leur appartenance à une classe déjà identifiée. Les spécialistes du machine learning testent habituellement différentes méthodes de classification pour identifier celle qui est la plus efficace compte-tenu du jeu de données d'entraînement, c'est-à-dire, celle qui génère un maximum de bonnes réponses pour un test réalisé avec un jeu de données en entrées qui sont déjà classées mais qui n'ont pas servi à l'entraînement du modèle.

Régression

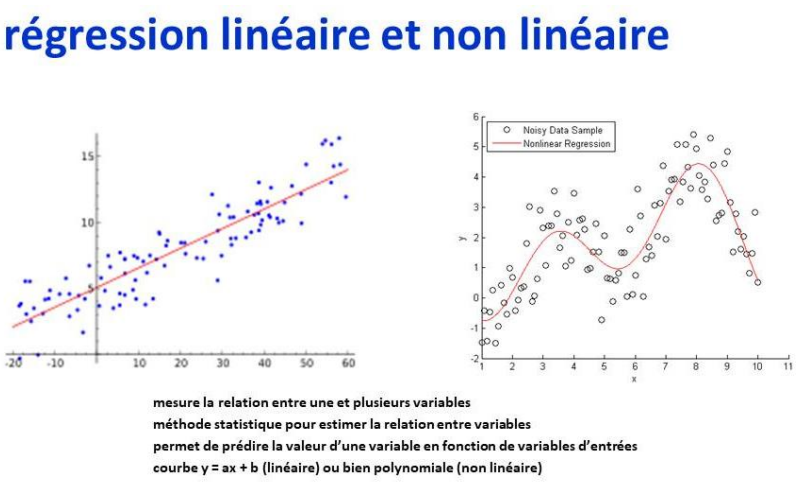
La régression permet de prédire une valeur numérique y en fonction d'une valeur x à partir d'un jeu d'entraînement constitué de paires de données (x, y).

[62] En pratique, on utilise diverses méthodes d'agrégation de résultats avec les adaboost, boosting, gradient boosting, XGBoost, bagging, LightGBM de Microsoft et autres random forest. C'est bien expliqué dans Ensemble Learning to Improve Machine Learning Results de Vadim Smolyakov, août 2017.

On peut par exemple prédire la valeur d'un bien immobilier ou d'une société en fonction de divers paramètres les décrivant. Les schémas *ci-dessous* qui illustrent ce concept utilisent uniquement une donnée en entrée et une en sortie. Dans la pratique, les régressions utilisent plusieurs paramètres en entrée.

Les jeux de donnée en entrée comprennent plusieurs variables (x, y, z, …). Il existe différentes formes de régression, notamment linéaire et non linéaire.

S'y ajoute aussi la notion d'overfitting et d'underfitting, qui décrit les méthodes de régression qui suivent plus ou moins de près les variations observées. Il faut éviter les deux et trouver le juste milieu ! C'est le travail des data scientists.

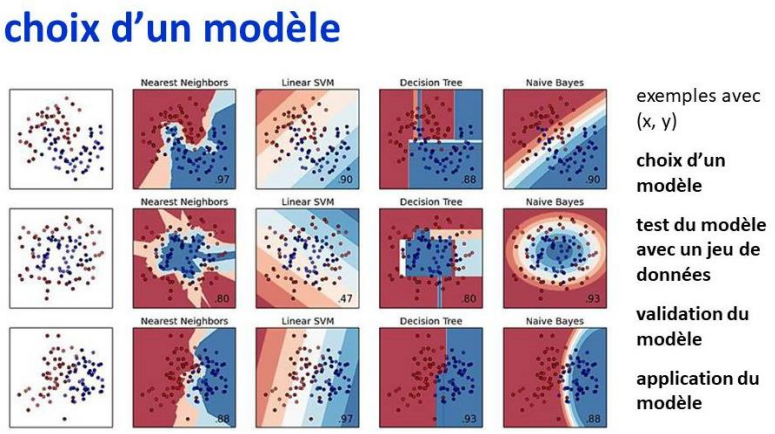


Les régressions peuvent être aussi réalisées avec des arbres de décision (CART), des modèles SVM, des réseaux de neurones, etc.

Clustering

Le clustering ou la segmentation automatique est une méthode d'apprentissage non supervisé qui permet à partir d'un jeu de données non labellisé d'identifier des groupes de données proches les unes des autres, les clusters de données. La technique la plus répandue est l'algorithme des k-moyennes (*k-means*).

Les méthodes de clustering permettent d'identifier les paramètres discriminants de ces différents segments. Elles servent ensuite à prévoir l'appartenance à un segment d'une nouvelle donnée entrée dans le système. Là encore, si le clustering peut être automatisé, en mode non supervisé, le choix du modèle de clustering ne l'est pas nécessairement pour autant sauf dans des outils avancés comme ceux de DataRobot et Prevision.io.



Le machine learning à base de réseaux de neurones permet de son côté de segmenter des données avec une répartition quasi-arbitraire alors que les méthodes élémentaires ci-dessus sont limitées de ce point de vue-là.

Réduction de la dimensionalité

La dimension des données devient de plus en plus grande à cause de la variété des big data. Un bon nombre d'algorithmes souffrent de la malédiction des grandes dimensions (« *curse of dimensionality* »).

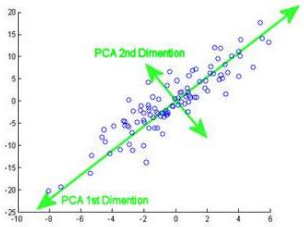
Il existe donc des techniques de nombreuses méthodes de réduction de dimension. Les les plus classiques consistent à plonger les données (on parle d'*embedding*) dans un espace de plus faible dimension, de façon à préserver certaines propriétés.

Par exemple, l'Analyse en Composantes Principales (ou PCA, « *Principal Component Analysis* ») est une projection linéaire sur un espace, dont la dimension est le nombre souhaité de dimensions final, qui préserve le mieux la variance ou la dispersion des données.

Dans le cas de l'Analyse Discriminante (Linear Discriminant Analysis : LDA), on projette les données linéairement sur un espace, mais en essayant de préserver au mieux la discrimination entre les classes.

On pourra ainsi, par exemple, identifier les paramètres d'une segmentation client ou leur combinaison qui sont les plus pertinents pour prédire un comportement donné (churn, achat, …) ou pour identifier les paramètres clés qui permettent la détection d'un virus informatique. Cela permet de simplifier les modèles et améliore les prédictions dans la suite des opérations.

la complexité du machine learning augmente avec le nombre de dimensions ou variables examinées
paramètres de segmentation marketing clients, phénotype patients, état de capteurs d'une machine pour maintenance prédictive, …



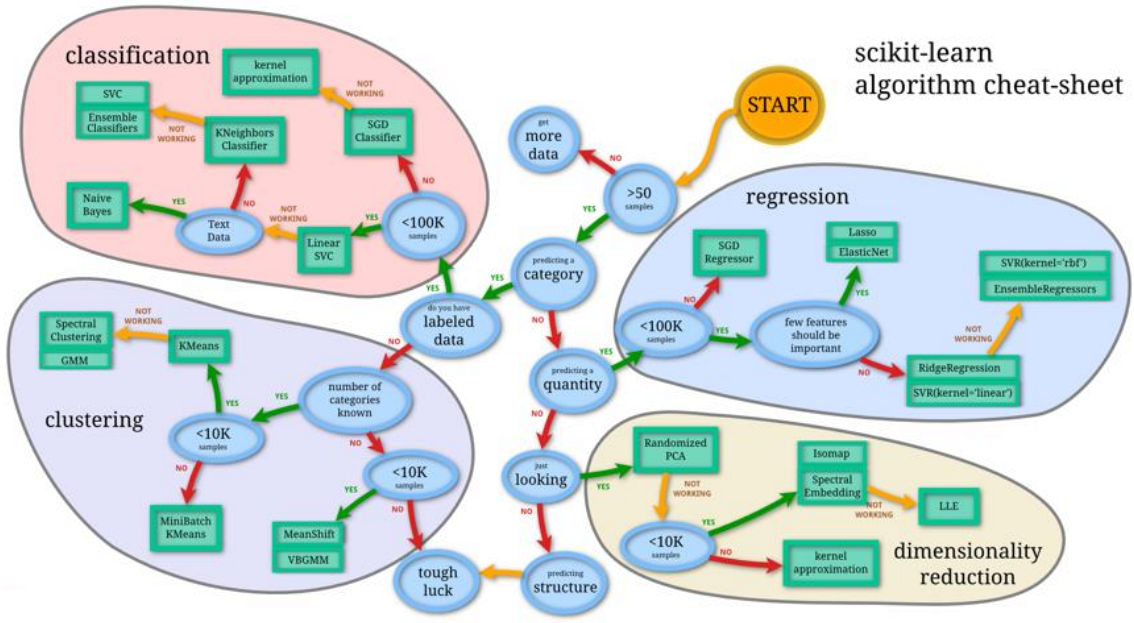
on réduit le nombre de dimension avec la méthode PCA (Principal Component Analysis)
elle permet d'identifier les variables clés discriminantes

La réduction du nombre de variables utilisées va aussi réduire la consommation de ressources machines. Mais attention, les variables discriminantes ou facteurs de corrélation ne sont pas forcément des facteurs de causalité. Ces derniers peuvent être externes aux variables analysées[63] !

Les techniques de réduction de dimension, et notamment la PCA, sont très largement utilisées dans le machine learning et le deep learning (qui a sa propre version de la PCA calculée par un réseau de neurones, dit auto-enccoder).

Outils du machine learning

Le machine learning nécessite d'abord de bien déterminer la typologie du problème à résoudre et des données disponibles. Le schéma ci-dessous originaire de scikit-learn est un exemple d'arbre de décision permetant de déterminer la méthode à utiliser en fonction du problème.



[63] C'est très bien expliqué dans cette tribune de Laurent Alexandre parue en novembre 2018 dans L'Express : Posséder un Picasso protégerait du cancer.

Il existe un très grand nombre d'outils de machine learning. Ils combinent plusieurs types de logiciels :

- Des **langages de programmation** comme Python, Java, C++ ou autres qui sont utilisés conjointement avec des bibliothèques de calcul spécialisées dans le machine learning. Il y a aussi le langage Julia associé aux bibliothèques JuliaStats qui permettent de créer des applications statistiques et de machine learning.
- Des **bibliothèques associées**, comme scikit-learn, d'origine française, qui permettent de développer les modèles d'apprentissage ou d'autoapprentissage et de les mettre ensuite en production. Ces outils tournent sur poste de travail et dans le cloud.

environnements de travail	bibliothèques
Apache Zeppelin	Scikit-Learn / Python
PyCharm	TensorFlow
Azure Machine Learning Studio	Mlpack / C++
Amazon Machine Learning	RapidMiner / Java
Google Cloud Machine Learning	Weka / Java
	Spark MLLib / Scala
	Torch / Lua
	JuliaStats / Julia

- Des **environnements de travail,** ou IDE pour Integrated Development Environment, qui permettent de paramétrer ses systèmes et de visualiser les résultats, souvent de manière graphique. Ils servent à tester différentes méthodes de classification, régression et clustering pour définir les modèles à appliquer. Ils peuvent aussi servir à piloter la mise en production des solutions retenues. Parmi eux, les solutions d'**IBM**, de **SAS**, de **Knime** (2008, USA, $20M), de **RapidMiner** (2007, USA, $36M)**,** les solutions de **Cognitive Scale** (2013, USA, $40M[64]) et **CrowdFlower** (2007, USA, $58M), le Data Science Workbench de **Cloudera** (2008, USA, $1B) et le Data Studio de **Dataiku** (2013, France, $43M). Il y a enfin **Matlab**. Cette catégorie comprend de nombreux autres acteurs tels que **Alteryx** (2010, USA, $163M), **Palantir** (2004, USA, $20B), **Predixion Software** (2009, USA, $37M), **Alpine Data Labs** (2011, USA, $25M) et **Lavastorm** (1999, USA, $55M).
- Des **outils d'automatisation** de la recherche de méthodes d'apprentissage comme **DataRobot** (2012, $225M, vidéo), le français **Prevision.io** (2016, vidéo et démonstration), le Coréen **Daria** (2015, $1,1M), **Solidware** (2014, Corée du Sud et créé par des français) propose Davinci Labs qui gère de manière intégrée tout le cycle de la création de solutions de machine learning. Enfin, Darwin de **SparkCognition** (2014, USA, $56,3M) est un outil équivalent. Ces outils récupèrent les données du client et permettent de tester diverses méthodes d'apprentissage relevant du machine learning pour trouver celles qui sont les plus pertinentes par rapport à un objectif à atteindre, de manière plus ou moins automatique selon les cas. Ils exécutent les tests de modèles en parallèle – parfois sur différentes ressources dans le cloud - pour prédire les valeurs d'une variable dans un tableau à partir d'un tableau d'entraînement[65].
- Des **outils destinés aux utilisateurs** pour leur permettre d'analyser leurs données et de produire des rapports graphiques pertinents en se passant théoriquement de data scientists. C'est ce que propose **Thoughtspot** (2012, USA, $150,7M) avec une solution qui utilise le machine learning pour identifier les besoins de présentation des données de l'utilisateur. Le montant levé par cette startup n'est pas très étonnant car elle vise un marché de volume, les utilisateurs et apporte beaucoup de valeur à ses clients d'entreprises.

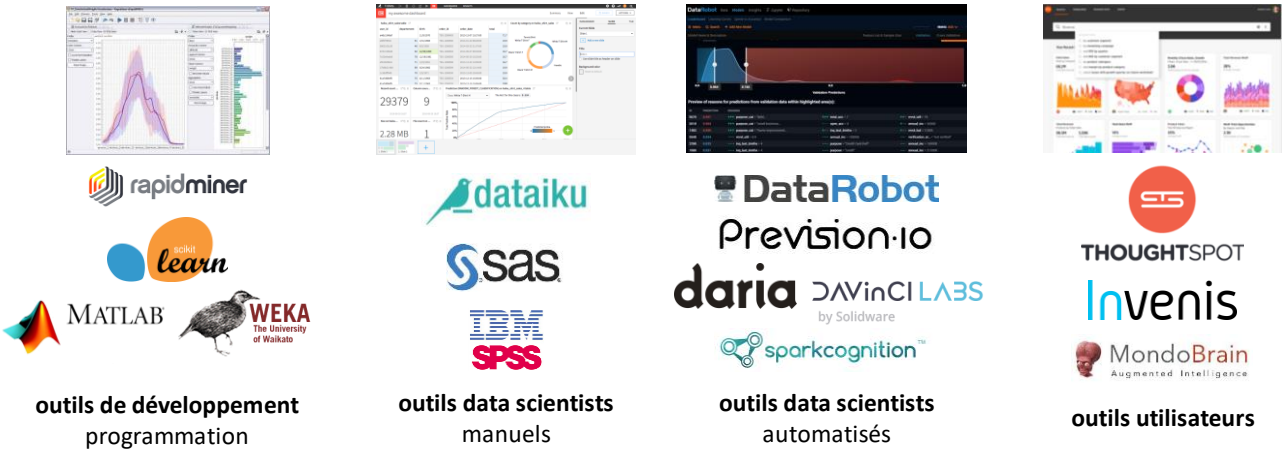
[64] Le marketing produit de Cognitive Scale est caricatural : il n'est franchement pas évidement de comprendre ce que réalise le produit. Celui de DataRobot est bien mieux réalisé et clair.

[65] Les outils permettent de se passer de programmation. Prevision.io créé un modèle prêt à l'emploi sans programmation et qui sera aussi exploitable par du code dans une application spécifique via une API en cloud. Voir ce descriptif précis du mode opératoire de prevision.io : Building a production-ready machine learning system with Prevision.io de Gerome Pistre, octobre 2017.

Invenis (2015, France) et **Mondobrain** (USA, $2,3M, vidéo) sont sur le même créneau, ce dernier générant automatiquement un dashboard multicritères à partir d'un jeu de données arbitraire[66]. **Tableau** (2003, USA, $15M, IPO en 2013) propose une sorte de Business Objects en cloud multiécrans et doté de fonctions collaboratives. Ils ont fait l'acquisition de la startup **Empirical Systems** (2015, USA, $2,5M) en juin 2018 pour ajouter à leur plateforme l'automatisation de la création de modèles de présentation de données. Bref, de faire du machine learning plus ou moins automatique. Il faut y ajouter **Dataiku** (2013, France, $45,7M) qui lançait en 2018 une nouvelle version de son Data Science Studio s'adressant aux utilisateurs qui peuvent ainsi prototyper, créer, déployer et gérer leurs modèles de données grâce à une interface graphique accessible et personnalisable. Des astuces telles que la fonction « Live Model Competition » permettent d'y comparer des modèles en temps réel sans attendre la fin des calculs de création du modèle.

- Des outils pour la création de solutions de machine learning pour les objets connectés, comme ceux de **Numericcal** (2016, USA). Ils optimisent le code généré pour tourner dans des systèmes contraints par la consommation d'énergie, la mémoire et la puissance disponibles.

Les compétences nécessaires pour créer des solutions de machine learning sont multiples. En amont, elles relèvent de la collecte et de l'organisation des données. C'est le big data. En son cœur, elle relève de la data science et des data scientists, qui exploitent ces données avec les logiciels du machine learning. Enfin, en aval subsistent des développeurs qui créent des solutions logicielles exploitables par les utilisateurs des entreprises ou le grand public.



Une bonne solution de machine learning doit être alimentée par des sources de données adaptées au problème à résoudre. Ces données doivent contenir suffisamment d'informations à valeur statistiques permettant de faire des régressions, segmentations ou prévisions. Leur bonne distribution spatiale dans l'univers du possible qui est étudié est encore plus importante que leur précision à l'échelle unitaire.

Réseaux de neurones

Les réseaux de neurones visent à reproduire approximativement par bio-mimétisme le fonctionnement des neurones biologiques avec des sous-ensembles matériels et logiciels capables de faire des calculs à partir de données en entrées et de générer un résultat en sortie. C'est une technique utilisée dans le machine learning et dans sa variante avancée du deep learning.

[66] Il s'agit presque toujours d'un tableau avec plusieurs colonnes de paramètres, la dernière colonne étant un « tag » déterminant d'un résultat, par exemple, un taux de défaut de fabrications de pièces, du churn client, ou tout autre comportement de machine ou d'individu.

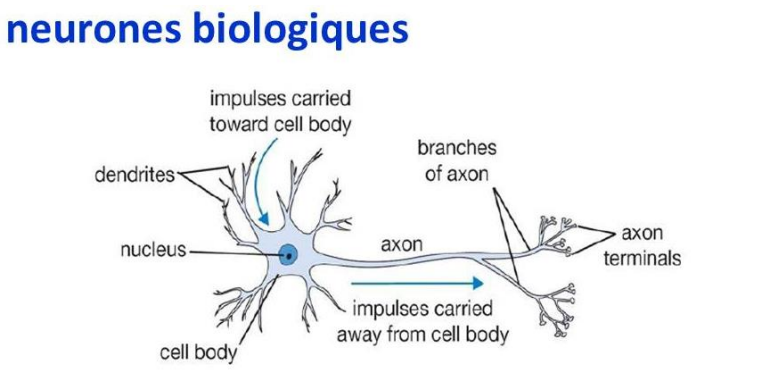
Les neurones artificiels

Le principe d'un neurone artificiel est de récupérer différentes variables numériques en entrée (x_1, …, x_n) associées à un poids (w_1, …, w_n) et à combiner ces valeurs pour générer une valeur en sortie. C'est un objet logiciel.

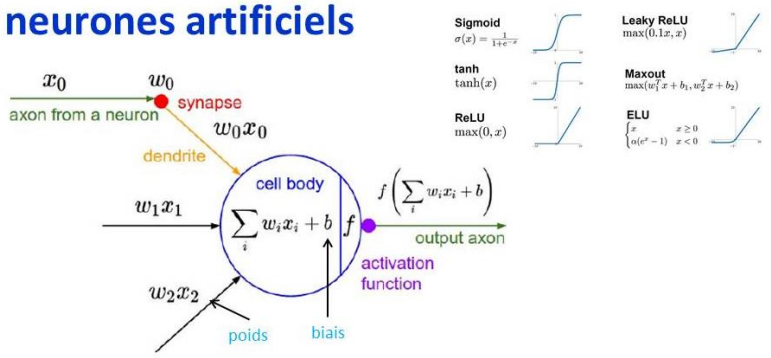
Le neurone artificiel moderne fait la somme des entrées multipliées par leur poids, additionne un biais (b qui qui permet de s'assurer que le résultat reste entre 0 et 1) et lui applique ensuite une **fonction d'activation** qui est en général une fonction non linéaire comme une sigmoïde qui génère une valeur comprise entre 0 et 1, générant une valeur % statistique facile à exploiter dans le reste du réseau de neurones.

La non linéarité de la fonction d'activation est une caractéristique clé des réseaux de neurones pour leur permettre de réaliser des fonctions complexes, et pas seulement linéaires.

Le procédé imite vaguement le fonctionnement d'un neurone biologique qui est largement plus complexe et dont le fonctionnement dépend d'un très grand nombre de paramètres biochimiques[67].



un neurone biologique opère une fonction d'activation encore non résolue en fonction des connexions avec d'autres neurones via dendrites / synapses / axones



additionne plusieurs plusieurs variables d'entrée avec des multiplicateurs ajustables (poids) et un biais, et y applique une fonction non linéaire (en général, sigmoide)

Un neurone isolé ne sert pas à grand-chose. Ils sont assemblés dans des réseaux de neurones. Un réseau de neurones de machine learning comprend souvent plusieurs couches de neurones. Les neurones d'une même couche ne sont généralement pas connectés entre eux contrairement aux neurones du cortex, histoire de créer des systèmes plus simples. Ils sont connectés aux neurones de la couche suivante.

On évite généralement des connexions circulaires entre neurones pour éviter de faire fonctionner le réseau en boucle lors de son apprentissage, sauf dans le cas des réseaux récurrents. C'est en tout cas vrai pour les réseaux de neurones convolutionnels que nous verrons plus loin.

Une couche cachée permet de générer une méthode de classification non linéaire complexe. On parle de deep learning lorsque le réseau de neurones comprend plus d'une couche cachée. C'est pour cela que le deep learning est considéré comme étant un sous-ensemble du machine learning[68].

---

[67] Un neurone du cortex cérébral est généralement relié par son axone à des milliers d'autres neurones via plusieurs synapses qui s'associent à une dendrite, une sorte d'excroissance de neurone. Il y a huit neurotransmetteurs différents qui font fonctionner les synapses. Et l'ensemble est régulé par l'expression de 6000 gènes différents dans les neurones et par des cellules gliales qui alimentent les neurones en énergie et qui régulent la production de neurotransmetteurs et la conductivité des axones via la myéline qui les entoure. Bref, c'est très compliqué ! Mais on en découvre tous les jours sur la complexité des neurones ! Voir Surprise! Neurons are Now More Complex than We Thought, de Carlos Perez, 2018. Et les milliers de microtubules qui constituent la structure des neurones pourraient elles-mêmes jouer un rôle clé dans la mémoire.

[68] Le deep learning est dénommé apprentissage profond en français mais j'utilise l'appellation anglaise dans ce document.

La « connaissance » du réseau de neurones est acquise via un processus d'apprentissage permettant d'ajuster le poids des interconnexions entre neurones pour que les objets en entrée du réseau de neurones soient reconnus en sortie, en général avec un tag descriptif, aussi appelé une classe, ou une valeur, comme le descriptif d'un objet pour une image en entrée. Il s'agit d'une connaissance purement probabiliste.

La connaissance d'un réseau de neurones n'est pas symbolique. Il ne sait pas donner de sens aux objets qu'il détecte ou aux calculs qu'il réalise ni expliquer les raisons de sa décision

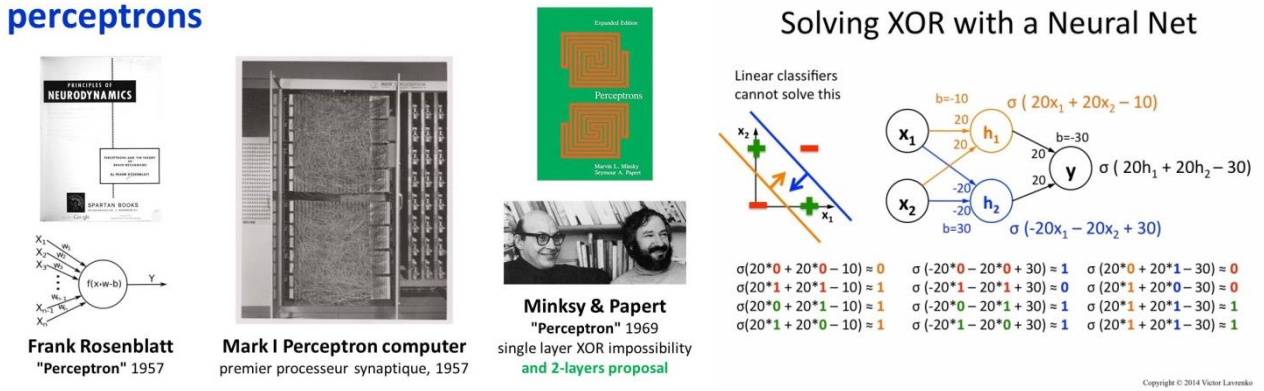
Les perceptrons

Le concept des réseaux de neurones a vu le jour en 1943 dans les travaux de **Warren McCullochs** et **Walter Pitts**. En 1949, **Donald Hebb** ajouta le principe de modulation des connexions entre neurones, permettant aux neurones de mémoriser de l'expérience.

Le premier réseau de neurones matériel fut créé par **Marvin Minsky** et **Dean Edmons** en 1950 alors qu'ils étaient étudiants à Harvard. Le SNARC simulait 40 neurones basiques avec 3000 lampes à tubes ! Et c'était avant le Summer Camp de Darmouth de 1956 !

Frank Rosenblatt, un collègue de Marvin Minsky, créa ensuite le concept du **perceptron** en 1957 qui était un neurone assez simple dans son principe avec une fonction de transfert binaire, générant un 0 ou un 1 en sortie.

Le premier perceptron était donc un réseau de neurones artificiels à une seule couche tournant sous forme de logiciel dans un **IBM 704**, le premier ordinateur du constructeur doté de mémoires à tores magnétiques. C'était un outil de classification linéaire utilisant un seul extracteur de caractéristique.



En 1969, **Marvin Minsky** publia avec **Seymour Papert** le livre **Perceptrons** qui critiquait les travaux de Frank Rosenblatt et sur un point très spécifique portant sur l'impossibilité de coder une porte logique XOR avec un perceptron. Une porte XOR détecte si les deux entrées binaires sont identiques : 0, 0 et 1, 1 deviennent 1 et 0,1 ou 1, 0 deviennent 0. Tout en proposant une solution de contournement associant deux couches de neurones pour mettre en œuvre la porte XOR. Le livre n'était donc pas si destructif que cela ! C'était même la voie vers les réseaux multi-couches qui, en effet, peuvent calculer un XOR. Mais il faudra attendre près de 20 ans pour qu'ils voient le jour.

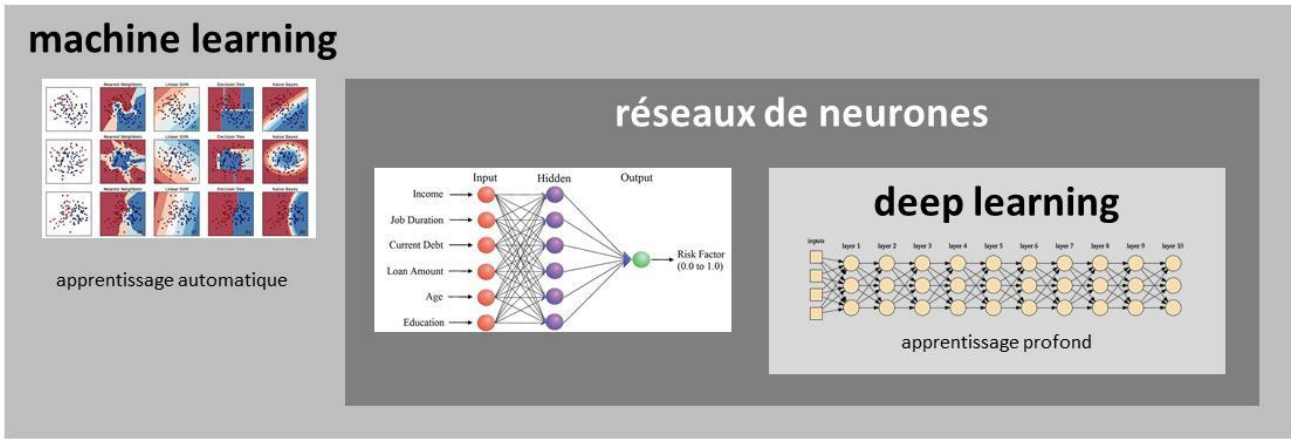
Les auteurs contribuèrent cependant à mettre un coup d'arrêt à ces développements, le coup de grâce arrivant avec le **rapport Lighthill** publié au Royaume Uni en 1973. Cela fit perdre un temps considérable à l'ensemble des recherches en IA, ce d'autant plus que les réseaux neuronaux sont devenus, depuis, un pan fondamental des progrès dans tous les étages de l'IA. Marvin Minsky reconnut toutefois son erreur d'appréciation dans les années 1980, après le décès de Frank Rosenblatt.

Du machine learning au deep learning

Les réseaux neuronaux ont connu ensuite un fort développement à partir de 2012 et dans leur mise en œuvre d'abord dans le machine learning puis avec le deep learning, qui exploite des réseaux de neurones avec un grand nombre de couches. C'est pour cela qu'ils sont profonds !

Dans le machine learning, les réseaux de neurones à une ou deux couches cachées permettent de créer des méthodes de classification d'objets plus sophistiquées.

De nombreuses méthodes d'organisation de réseaux de neurones sophistiqués sont ensuite apparues pour permettre la reconnaissance de la parole et d'images. Elles sont évoquées dans la partie sur le deep learning.



méthodes simples
de classification, régression et segmentation

imitent le fonctionnement des neurones biologiques

les neurones d'une même couche sont **connectées à la couche suivante** pour simplifier l'entrainement

une **couche cachée** permet de gérer une méthode de classification non linéaire

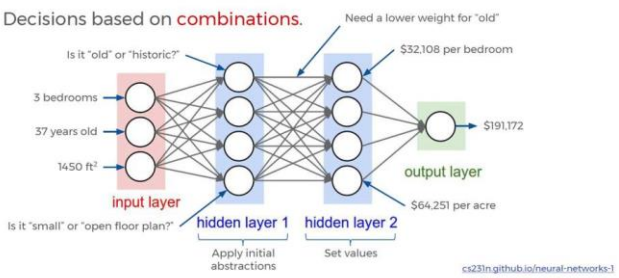
on parle de **deep learning** lorsque le réseau de neurones comprend **plusieurs couches cachées**

principaux modèles de DL :
réseaux convolutionnels (spatiaux)
réseaux à mémoire (temporels)
génératifs (interpolation de contenus)

Enfin, citons les réseaux de neurones multi-modes qui exploitent des sources d'informations complémentaires, classiquement, de l'audio et de la vidéo, pour améliorer la qualité de la captation. L'audio d'une vidéo permet par exemple d'améliorer la capacité à tagger le contenu de la vidéo. Cela peut aller jusqu'à lire sur les lèvres pour améliorer la reconnaissance de la parole.

L'imagerie 2D complétée par des informations de profondeur améliorera la capacité de détection d'objets complexes. La vidéo d'un visage permettra d'améliorer la captation de la parole par l'équivalent numérique de la lecture sur les lèvres.

réseau de neurone et évaluation d'un prix



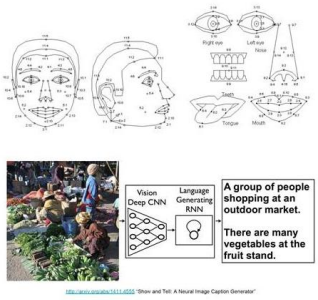
réseaux de neurones multi-modes

associent ≠ types de données

image/vidéo + texte => descrition du contenu

vidéo + audio => reconnaissance de la parole

couleur + profondeur => reconnaissance d'objet et navigation



Un exemple classique « de dans les livres » de réseau de neurones simple est celui de l'estimation du prix d'un appartement en fonction de quelques critères clés discriminants comme sa surface, son âge et son ancienneté.

Avec quelques paramètres numériques de ce type, un tel réseau peut se contenter de n'avoir que quelques couches, deux dans l'exemple. En sortie de réseau, il génèrera une estimation du prix de l'appartement[69].

Ce sont des réseaux multi-couches dits **feed forward** : on les alimente en amont avec des données qui rentrent dans les neurones de la première couche puis passent aux neurones de la couche suivante via leurs synapses, ainsi de suite jusqu'à la dernière couche qui donne une réponse.

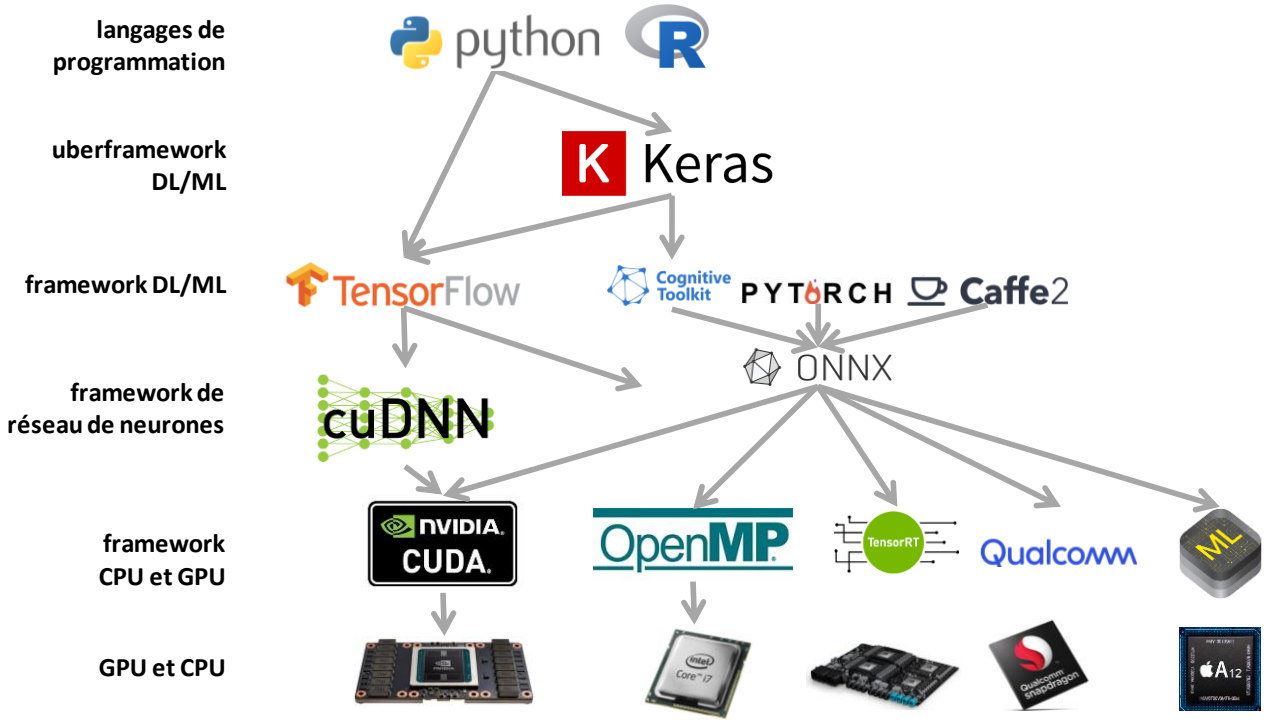
Sur les schémas, l'information circule de gauche à droite pendant l'exécution du réseau de neurone. On appelle aussi cela une inférence.

Comment entraîne-t-on un réseau de neurones, à savoir, comment ajuste-t-on le poids de chacune des synapses de chaque neurone du réseau ? Nous répondrons à cette question dans la partie consacrée au deep learning et à la rétropropagation d'erreurs.

Programmation de réseaux de neurones

D'un point de vue pratique, la programmation de réseaux de neurones s'appuie sur des bibliothèques logicielles spécialisées comme **cuDNN**, **MKL** ou **OpenNN**. On peu aussi citer **Synaptic** qui est une bibliothèque utilisable avec node.js dans un navigateur en JavaScript.

Ces bibliothèques de réseaux de neurones sont souvent exploitées elles-mêmes par des bibliothèques de machine learning ou de deep learning, comme **TensorFlow**, qui masquent la complexité du pilotage de réseaux de neurones à bas niveau et permettent par exemple de définir les modèles de réseaux de neurones convolutionnels de reconnaissance d'images et de les entraîner.



C'est illustré dans le schéma *ci-dessus* qui empile les couches utilisées dans le développement de solutions d'IA avec un développement comprenant un framework d'asbtraction élevé utilisant un framework, comme le framework **Keras** qui se situe au-dessus de TensorFlow, puis une bibliothèque de réseau de neurones, suivie d'une bibliothèque de pilotage de GPU comme CUDA chez Nvidia, et enfin, un GPU ou un CPU au niveau matériel.

---

[69] Une solution voisine de cet exemple académique semble être maintenant opérationnelle sur le site SeLoger. Voir Dans les coulisses du nouvel outil d'estimation de prix de SeLoger de Justine Gay, octobre 2018.

Les frameworks **CNTK** de Microsoft**, PyTorch** et **Caffe 2** de Facebook supportent pour leur part un format de description de modèle intermédiaire ONNX qui supporte de son côté les principaux frameworks de CPU et GPU du marché.

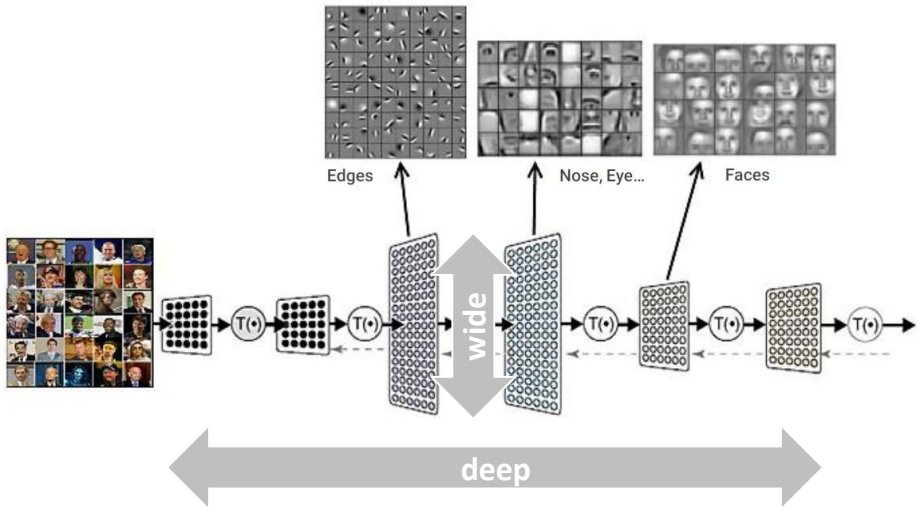
TensorRT est le framework d'exécution de modèles entraînés de Nvidia. **Qualcomm** a le sien pour ses Snapdragon, le Qualcomm Neural Processing SDK for AI. **CoreML** est de son côté lié aux plateformes Apple.

Deep learning

Le deep learning est un sous-ensemble des techniques de machine learning à base de réseaux de neurones qui s'appuient sur des réseaux de neurones à plusieurs couches dites cachées.

Celles-ci permettent par exemple de décomposer de manière hiérarchique le contenu d'une donnée complexe comme de la voix ou une image pour la classifier ensuite : identifier des mots pour la voix ou associer des tags descriptifs à des images.

C'est le principe de l'une des grandes catégories de réseaux de neurones de deep learning, les réseaux convolutionnels ou convolutifs (schéma *ci-dessous*). Un réseau peut être profond mais aussi large si le nombre de neurones est élevé dans chaque couche. Le deep learning remplace les méthodes antérieures du machine learning dites à base de « handcraft features » qui consistaient à définir à la main les éléments à rechercher dans les objets (formes dans les images, etc). Dans le deep learning, notamment pour la détection d'images, le réseau de neurones découvre tout seul ces composantes avec des niveaux d'abstraction évoluant de bas en haut et de couche en couche.



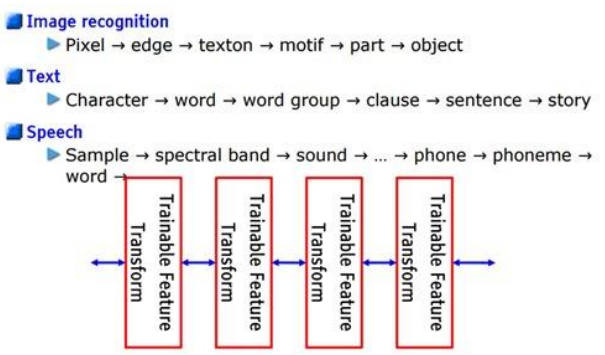
chaque niveau d'un réseau de neurones profond réalise une fonction de classification d'un type d'objet en un type d'objet supérieur
(pixels => forme => nez => visage => individu)

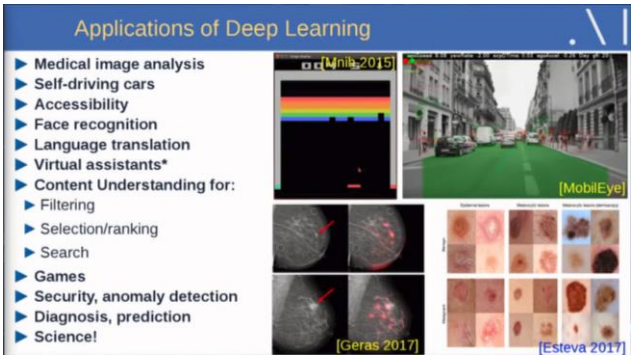
un réseau de neurones profond de type convolutionnel comprend plusieurs couches "cachées" qui transforment les données en entrée en données ayant un niveau d'abstraction supérieur

Le deep learning sert le plus souvent au traitement du langage, de la parole, du bruit, de l'écriture et des images. Il a d'autres usages dans les outils d'aide à la décision, dans les jeux tels que le Go avec AlphaGo et même dans l'exploitation de données structurées, dans la cybersécurité et d'une manière générale dans la recherche scientifique[70] comme en génomique.

Tous les secteurs d'activité peuvent faire appel au deep learning comme nous l'explorerons dans la grande partie correspondant aux applications métiers de l'IA.

[70] Le slide de droite est issu de la conférence de Yann LeCun à l'USI à Paris en juin 2018 (vidéo). Celui de gauche provient également de Yann LeCun, mais de sa conférence inaugurale au Collège de France en 2016 (lien).



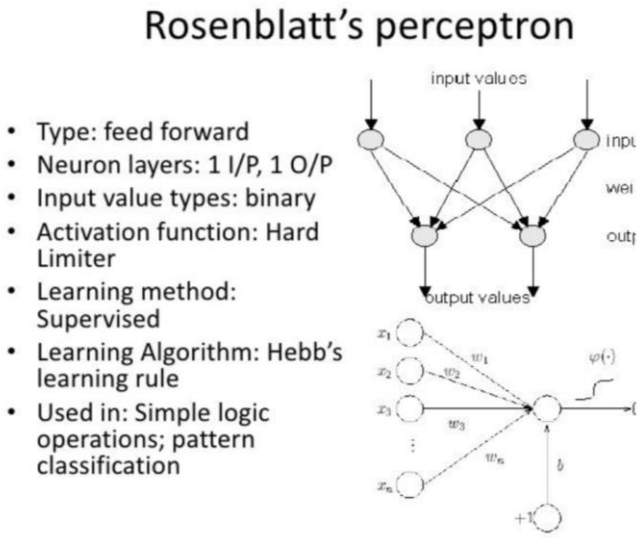


Le deep learning permet aussi de générer des contenus artificiels, extrapolés à partir de contenus réels, notamment des images, que nous verrons aussi, et qui s'appuient sur des modèles génératifs et « adversariaux » (ou GAN, pour Generative Adversarial Networks).

Par contre, le deep learning n'est pas la panacée. Il s'appuie sur des modèles probabilistes comme son papa le machine learning. Il n'est pour l'instant pas adapté au raisonnement, à la génération de sens commun, à la création de robots intelligents et agiles et d'une manière générale à la création d'IA générales (AGI, dont nous parlerons plus tard). Cela pourrait cependant changer un jour.

Evolutions du deep learning

Les outils de deep learning s'appuient sur différentes variantes de réseaux de neurones pour leur mise en œuvre pratique. Leur histoire remonte aux perceptrons de **Franck Rosenblatt** de 1957.



perceptrons 1957
multi-layered perceptron 1969
back propagation 1974
recurrent neural networks 1982
neocognitrons 1983
error propagation 1986
restricted boltzmann machine 1986
time delay neural networks 1989
forward propagation 199X
convolutional neural networks 1998
deep belief networks 2006
stacked autoencoders 2007
google imagenet 2012

L'histoire du deep learning a véritablement démarré plus de 20 ans plus tard, dans les années 1980. Il a cependant fallu attendre 1995 pour que l'on puisse les mettre en œuvre en pratique, sans doute, grâce aux progrès matériels, à la loi de Moore mais aussi aux progrès conceptuels, notamment aux travaux d'Alexander Weibel en 1989, Yann LeCun en 1988 et 1998 et à Geoff Hinton, particulièrement à partir de 1986. Le premier est le père des réseaux TDNN de reconnaissance de phonèmes, le second des réseaux convolutifs à rétropropagation d'erreurs tandis que le dernier est considéré comme étant le père de nombreux concepts de base du deep learning.

Il est de bon ton de déclarer que les chercheurs n'ont pas produit grand chose depuis et que le deep learning doit tout aux progrès du matériel et à l'abondance de données pour entraîner les systèmes.

Quand on y regarde de plus près, on se rend compte qu'au contraire, les chercheurs n'ont pas cessé de faire avancer le domaine[71]. Et d'année en année, des progrès conceptuels et pratiques font avancer les réseaux de neurones et le deep learning, ne serait-ce qu'avec les réseaux génératifs.

Le champ du deep learning est en perpétuelle évolution. On voit fleurir régulièrement de nouveaux types de réseaux de neurones, que ce soit pour la reconnaissance d'images (identification d'objets, segmentation graphique d'objets, labellisation d'objets, temps réel) ou dans le traitement du langage (traduction, questions/réponses, représentation des connaissances). Chaque année, un nouveau réseau de neurones rend obsolète ceux de l'année précédente. C'est un monde de remise en cause permanente de l'état de l'art.

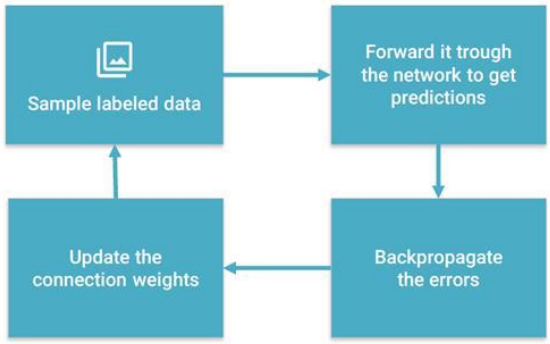
C'est ce que nous allons voir dans ce qui suit. Ces avancées du deep learning sont étalées sur plusieurs décennies et sont continues. Elles sont évoquées ici de manière chronologique selon leur date d'apparition.

Rétropropagation d'erreurs (1969)

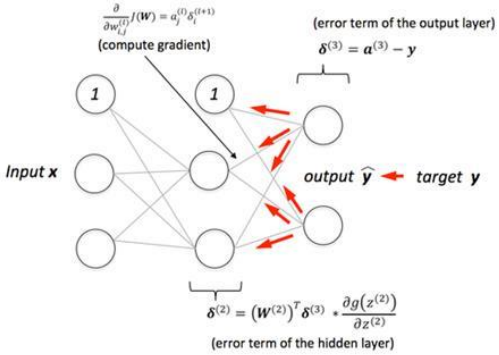
Un réseau de neurones a besoin d'être entraîné, à savoir que le poids des synapses a besoin d'être ajusté pour que le réseau de neurones génère de bons résultats. Or ces paramètres sont très nombreux dans un réseau de neurones de deep learning. Ils peuvent être plusieurs milliards !

Comment fait-on donc pour définir le poids de ces synapses ? La méthode la plus courante consiste à utiliser la rétropropagation du gradient. Elle fonctionne couche par couche en partant du résultat et en ajustant le poids des neurones pour permettre au réseau d'identifier les objets de la base d'entraînement fournis en entrée.

Cette rétropropagation fonctionne en ajustant un par un les poids des neurones de chaque couche et en scannant un par un les objets du jeu de test pour optimiser le taux de reconnaissance, en minimisant ce que l'on appelle la « fonction d'erreur », soit la différence entre ce que génère le réseau pendant sa phase d'entraînement et la bonne réponse dont on dispose déjà dans la base d'entraînement.



l'apprentissage supervisé avec rétro-propagation d'erreurs mesure la différence entre prédictions et valeurs souhaitées, celle-ci servant à modifier les poids des neurones pour améliorer les prédictions



pour chaque objet testé, l'erreur est répercutées dans les neurones amont en modifiant les poids des synapses au prorata de leur poids respectif

L'apprentissage des réseaux de neurones est généralement supervisé et automatique ! Supervisé car il utilise des labels descriptifs des objets d'une base de référence et automatique car les poids synaptiques des neurones sont ajustés automatiquement grâce à ces méthodes de rétropropagation programmées dans le système d'entraînement.

Les évolutions des méthodes de rétropropagation créées par la suite visaient surtout à économiser du temps machine car l'opération est très fastidieuse puisqu'elle doit être répétée pour chaque neurone du réseau et pour chaque objet de la base de référence. Cela donne une combinatoire très élevée !

---

[71] Voir par exemple Beyond backpropagation : can we go deeper than deep learning ? de Marina Yao, novembre 2017.

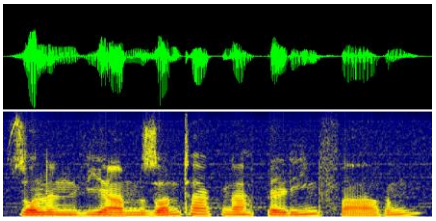
La méthode est perfectionnée en 1986 par David Rumelhart, Geoff Hinton et Ronald Williams dans Learning representations by back-propagating errors, alors qu'au même moment, Yann LeCun publiait une version de l'algorithme dans sa thèse (en 1987). La plus couramment utilisée aujourd'hui est la **descente stochastique de gradient** (ou SGD pour stochastic gradient descent en anglais), vue un peu plus loin, et qui permet d'améliorer la vitesse de convergence des réseaux lors de leur entraînement.

La *backprop* a ceci de particulier qu'elle ne relève pas a priori du biomimétisme. Ou tout du moins, on n'en sait rien puisque l'on ne connaît pas les mécanismes biologiques d'entraînement du cerveau.

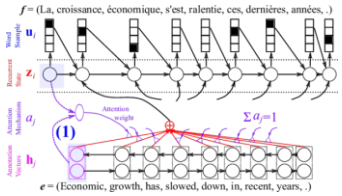
Réseaux de neurones récurrents et à mémoire (1982 puis 1993)

Ces RNN (Recurrent Neural Networks) permettent d'analyser des informations évoluant dans le temps comme la voix au niveau des phonèmes et le langage au niveau de l'assemblage des mots. Ils sont en effet très utilisés dans les systèmes de reconnaissance de la parole, pour la traduction automatique et la reconnaissance de l'écriture manuscrite.

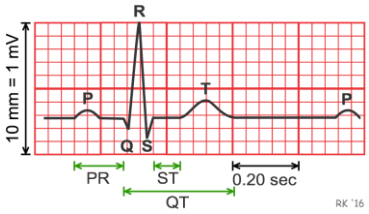
Ils peuvent aussi analyser des signaux comme le bruit ou les vibrations de machines pour y détecter des anomalies, dans le cadre de maintenance préventive, aux prévisions de cours d'action, de la consommation d'énergie ou d'eau pour les utilities, à l'analyse d'électro-cardiogrammes[72] et même à la détection des exoplanètes par la méthode des transits[73].



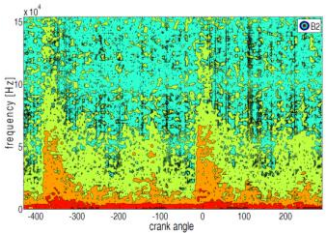
reconnaissance de la parole
(+ CNN)



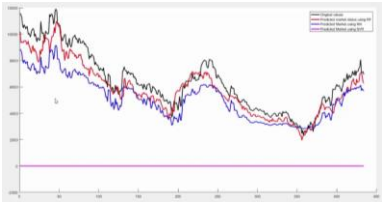
traduction automatique



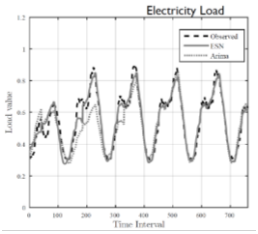
analyse d'ECG
(+CNN)



maintenance préventive



prévisions boursières



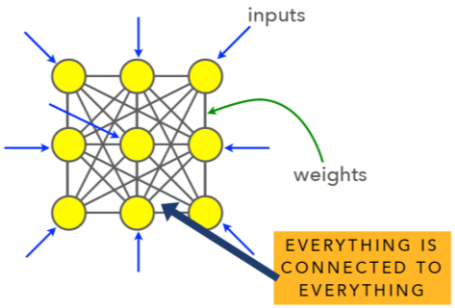
consommation

Les réseaux de neurones à mémoire ont bien évolué dans le temps avec différentes déclinaisons qui se sont inspirées les unes les autres :

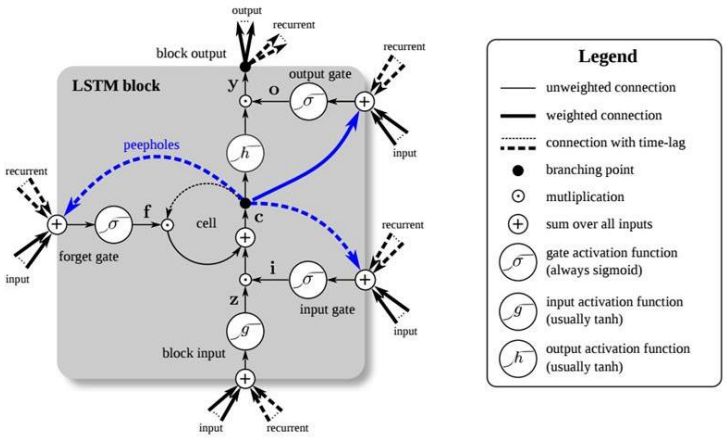
[72] Concommitamment avec d'autres méthodes comme les CNN, réseaux convolutionnels.

[73] Que j'ai eu l'occasion d'expliquer ici : http://www.oezratty.net/wordpress/2017/astronomie-entrepreneuriat-exoplanetes/.

- **Hopfield Network** (1982) est un des premiers réseaux à mémoire qui imite le fonctionnement de l'hypocampe du cerveau. Son créateur, un physicien, a relancé l'intérêt des réseaux de neurones qui étaient en rade depuis 1969. Le réseau de Hopfield permet de gérer une mémoire associative bidirectionnelle. Cela se retrouve dans sa matrice de connexions qui est symétrique et nulle dans sa diagonale. Tous les neurones du réseau sont connectés aux autres. La technique est cependant limitée en termes de stockage[74].



- **BPTT** (1987), BackProp Through Time, une méthode d'entraînement de réseaux de neurones récurrents.
- **RTRL** (1989), Real Time Recurrent Learning, une variante de réseaux de neurones récurrents.
- **Simple Recurrent Network (SRN)** ou **Ellman Network** (1990) gère une couche cachée de contexte. Il s'entraîne par rétropropagation[75].
- **LSTM** (1997), Long Short Term Memory, qui savent gérer le contexte dans lequel les contenus apparaissent [76] et sont très utilisés pour le traitement du langage et la traduction automatique. Ce sont des réseaux en quelque sorte récursifs. Ils ont été créés par l'Allemand Sepp Hochreiter et le Suisse Jurgen Schmidhuber. Ils sont encore aujourd'hui une base clé du traitement du langage.



- **Reservoir Computing** (2005) utilise un « réservoir » de neurones constituant un système complexe et dynamique qui n'exploite pas l'approche classique du feed forward mais plutôt quelque chose qui pourrait s'apparenter à la notion de ligne à retard, comme dans l'audio[77]. On l'entraîne avec une fonction linéaire qui consolide les valeurs de l'ensemble du réservoir dans dans une couche de classification.

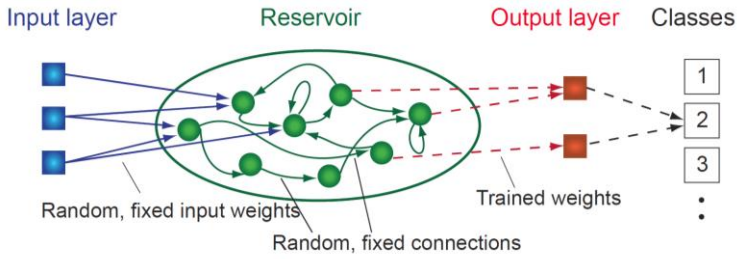


Fig. 1.8: Classical reservoir computing scheme. The input is coupled into the reservoir via a randomly connected input layer to the N nodes in the reservoir. The connections between reservoir nodes are randomly chosen and kept fixed, that is, the reservoir is left untrained. The reservoir' s transient dynamical response is read out by an output layer, which are linear weighted sums of the reservoir node states. Figure taken from Appeltant *et al.* [17].

[74] Voir Hopfield Networks (29 slides).

[75] Voir The Simple Recurrent Network: A Simple Model that Captures the Structure in Sequences qui décrit bien la logique des SRN.

[76] Les LSTM ont été conceptualisés par Sepp Hochreiter et Jürgen Schmidhuber dans Long short-term memory, en 1997. Ce dernier est le créateur de la startup suisse **nNaisense**, qui ambitionne de créer une AGI (Artificial General Intelligence).

[77] Voir la présentation Introduction to Reservoir Computing, de Helmut Hauser, 2013 (282 slides) et la source du schéma Reservoir computing based on delay-dynamical systems (160 pages).

Cette couche de classification ressemble à celle de la fin d'un réseau convolutionnel (que nous verrons plus tard). Ce genre de réseau est adapté aux systèmes où règne le chaos. Il est utilisé dans les prévisions financières, dans le traitement du son, dans la détection d'épilepsies, dans la localisation de robots, dans la gestion de structures grammaticales. Il existe trois variantes de Reservoir Networks : Liquid State Machine, Echo State Network et Backpropagation-Decorrelation learning rule.

- **GRU** (2014), Gated Recurrent Units[78] est une des variantes plus simples des LSTM et est très utilisée. Elle évite notamment l'apparition du problème de diminution des gradients (« vanishing gradient problem ») qui empêche l'entraînement du réseau de neurones du fait de gradients de propagation d'erreurs trop faibles.
- **BLSTM** (2015), Bidirectionnal Long Short Term Memory, sont des LSTM bidirectionnels.
- **Tree-LSTM** (2015) sont des LSTM organisés en arbres et non linéairement. Ils peuvent notamment servir à l'analyse de sentiment dans des textes[79].
- **Stacked RNN** (2015) sont des RNN empilés[80] comme leur nom l'indique.
- **MANN** (2015), des Memory-Augmented Neural Networks qui permettent un entraînement plus rapide des réseaux à mémoire et avec un jeu de tests plus limité.
- **Transformers** (2017) sont des réseaux de neurones convolutionnels qui utilisent un mécanisme de gestion de l'attention. Ils sont adaptés à la traduction et plus rapides à entraîner que les LSTM et autres variantes de réseaux récurrents[81].

Je vous épargne les détails de toutes ces variantes de réseaux récurrents, ce d'autant plus que je n'ai pas encore très bien compris leur fonctionnement dans le détail.

Ils sont difficiles à vulgariser[82], bien plus que les réseaux de neurones convolutionnels que nous verrons un peu plus loin, et qui permettent d'analyser le contenu d'images. Ces réseaux récurrents sont d'ailleurs souvent combinés entre eux de manière plus ou moins empirique.

Ces réseaux transforment généralement les mots et phrases en vecteurs, des objets mathématiques triturés pour être comparés les uns aux autres, classifiés, modifiés et transformés. Ils permettent surtout de tenir compte du contexte dans lequel les objets comme des mots sont détectés pour analyser le sens d'une phrase. L'un des points clés de ces réseaux est leur capacité à mémoriser des contextes[83].

C'est un domaine d'amélioration encore plus intense que dans les réseaux de neurones convolutionnels. Avec à la clé des solutions de plus en plus performantes pour la reconnaissance de la parole, la traduction automatique et les agents conversationnels réellement intelligents.

Machines de Boltzmann restreintes (1986)

Les machines de Boltzmann restreintes utilisent une seule couche de neurones source et cible.

---

[78] Les GRU ont été créés par Junyoung Chung en 2014. Voir ce papier de Junyoung Chung, Caglar Gulcehre, KyungHyun Cho et Yoshua Bengio Empirical evaluation of gated recurrent neural networks on sequence modeling qui compare les GRU aux LSTM.

[79] Voir Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks de Kai Sheng Tai et Christopher Manning de Stanford et Richard Socher de Metamind, une startup acquise par Salesforcee, 2015 (11 pages).

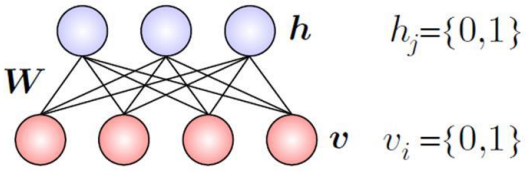
[80] Voir Meta-Learning with Memory-Augmented Neural Networks, 2016 (9 pages).

[81] Voir Attention Is All You Need, 2017 (11 pages), Universal Transformers, juillet 2018 (19 pages) et Transformer: A Novel Neural Network Architecture for Language Understanding de Jakob Uszkoreit, 2017.

[82] Voir le cours de Stanford Recurrent Neural Networks de Fei-Fei Li, Justin Johnson et Serena Yeung, mai 2018 (107 slides).

[83] Voir la conférence de Rob Fergus au Collège de France en avril 2016 dans le cadre de la chaire de Yann LeCun.

Il n'y a pas de connexions entre les neurones d'une même couche. C'est le modèle le plus simple de réseau de neurones qui est ensuite exploité dans d'autres assemblages, comme les Deep Belief Networks (DBN) créés en 2006 par Geoff Hinton.



Time Delay Neural Networks (1989)

Créés par l'Allemand Alexander Weibel en 1989, les TDNN permettent notamment de reconnaître des phonèmes dans la parole sans avoir à les positionner explicitement dans le temps[84]. Les TDNN peuvent être considérés comme des précurseurs des réseaux convolutionnels. Ces derniers reconnaissaient au départ des images tandis que les TDNN reconnaissaient des sons qui sont aussi des données bidimentionnelle (fréquence, temps). La différence entre les TDNN et les ConvNets se situe surtout dans l'ajout de couches de pooling dans ces derniers. Nous verrons cela plus loin.

Les TDNN ont été utilisés en France dans l'équipe de Françoise Soulié Fogelman pour la parole[85] et les images[86], et au LIMSI pour la parole avec Laurence Devillers en 1992.

Réseaux de neurones convolutionnels (1989 puis 1998)

En 1998, le français Yann LeCun[87], qui avait quitté l'équipe de Soulié Fogelman après sa thèse en 1987, pour rejoindre celle de Geoff Hinton à Toronto, puis celle de Larry Jackel aux Bells Labs en 1988, étend le modèle des TDNN à la reconnaissance de caractères puis à la reconnaissance d'images[88].

En 2012, le réseau AlexNet de Alex Krizhevsky et Geoffrey Hinton écrase les autres techniques à la compétition ILSVRC (ImageNet Large-Scale Visual Recognition Challenge), établissant la supériorité des réseaux de convolution pour la reconnaissance d'images face aux méthodes traditionnelles du machine learning.

Ce sont des outils qui servent principalement à réaliser de la classification d'objets, comme pour associer une image à une classe d'objets (chat, bateau, avion, …) ou un phonème vocal à son identifiant (/a/, /u/, ..).

Les premiers CNN de production ont été déployés en 1995 pour la reconnaissance des chèques, via une solution de NCR.

---

[84] Voir Review of TDNN (Time Delay Neural Network) – Architecture for Speech Recognition de Masahide Sugiyamat, Hidehumi Sazoait et Alexander Waibel qui font le point en 1991 de l'état de l'art des TNDD (4 pages).

[85] Voir Experiments with time delay networks and dynamic time warping for speaker independent isolated digits recognition de Léon Bottou et al, 1989 (4 pages).

[86] Voir Scene segmentation using multiresolution analysis and MLP de Emmanuel Viennet et Françoise Fogelman Soulié, 1992.

[87] Yann LeCun s'était inspiré des travaux de Kunihiko Fukushima, un chercheur de la NHK, et de ses réseaux de neurones multicouches Neocognitron. Voir Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition, 1987. Et le papier de 1998 Gradient-based learning applied to document recognition dont lesauteurs sont Yann LeCun, Léon Bottou, Yoshua Bengio et Patrick Haffner (46 pages).

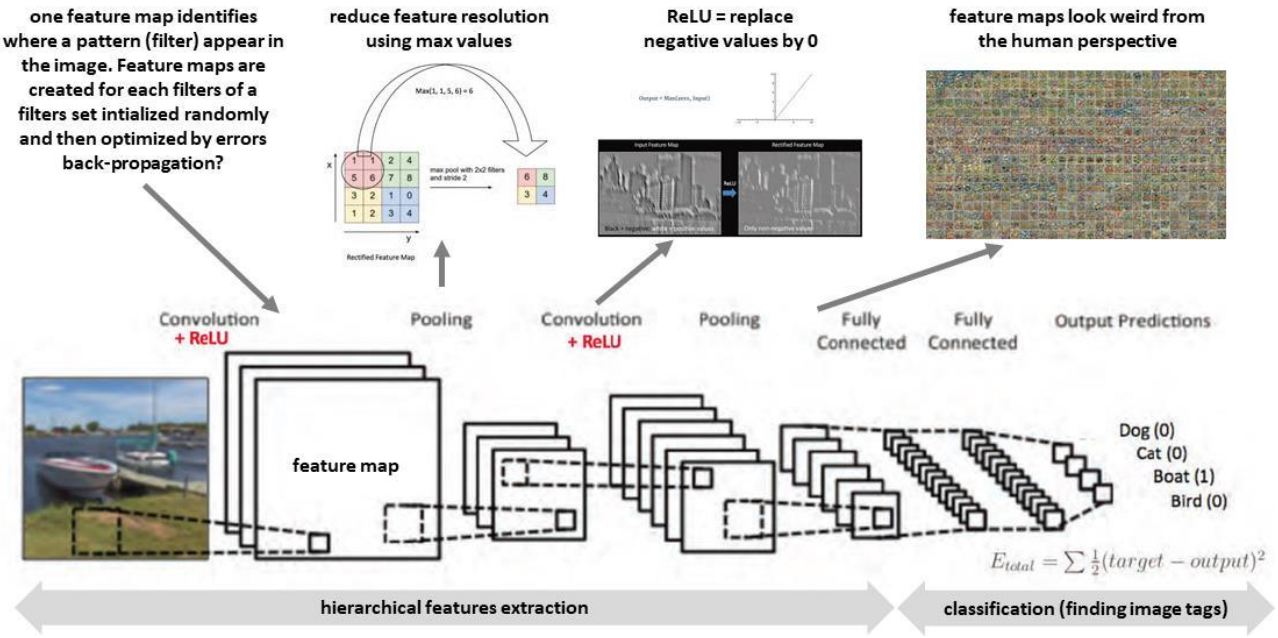
[88] Voir cette bonne explication en trois parties : A Beginner's Guide To Understanding Convolutional Neural Networks de Adit Deshpande (un étudiant aux USA), partie 1, partie 2 et partie 3, 2016.

Les CNN, appelés aussi ConvNets (convolutional neuron networks), utilisent plusieurs techniques enchaînées les unes avec les autres avec notamment des filtres et des feature maps qui consistent à identifier des formes dans les images, avec des niveaux d'abstraction allant du plus petit au plus grand. Depuis, la technique a évolué avec de nombreuses variantes de plus en plus performantes[89].

Une **feature map** est une matrice de pixels qui cartographie de l'apparition d'un filtre donné dans l'image analysée. Un ConvNet utilise un jeu de plusieurs filtres initialisé aléatoirement, sauf pour la première couche de convolution qui est généralement initialisée avec des filtres classiques décrivant des transitions horizontales, verticales et diagonales dans les images, une technique utilisée dans les méthodes d'analyse d'images antérieures aux réseaux convolutionnels. Les filtres sont des matrices de quelques pixels de côté, en général 3x3 ou 4x4[90].

Ils sont ensuite affinés par rétropropagation d'erreurs de l'ensemble du réseau, un mécanisme qui est appliqué pour toutes les images d'un jeu d'entraînement qui peut comprendre des millions d'images[91] et même 3,5 milliards chez Facebook labellisées avec 15 000 hashtags différents[92].

Cet entraînement est très consommateur de ressources machine et aussi d'accès à la mémoire mais bien plus efficace qu'un simple réseau de neurones multicouches. Cela vient du fait que le réseau comprend moins de paramètres. Ce nombre de paramètres est approximativement égal à la somme de l'information des nombreux filtres de chaque convolution et des poids des synapses des couches terminales du réseau.



[89] Voir l'excellente série de papiers A Beginner's Guide To Understanding Convolutional Neural Networks, A Beginner's Guide To Understanding Convolutional Neural Networks Part 2 et The 9 Deep Learning Papers You Need To Know About (Understanding CNNs Part 3), 2016 qui décrit les différentes générations de CNN qui s'étaient alors succédées : AlexNet (2012, avec cinq couches de convolution, entraîné sur 15 millions d'images), ZF Net (2013), VGGNet (2014, qui utilise des petits filtres de 3x3 pixels), GoogleLeNet (2015, qui utilise des filtres de taille différente en parallèle et une centaine de couches en tout), Microsoft ResNet (2015, avec 152 couches), les R-CNN (2013-2015), puis les GAN (2014). Voir aussi cette vidéo d'Andrej Karpathy, Deep Learning for Computer Vision, septembre 2016 (1h25)

[90] On retrouve cette taille de matrices dans les processeurs neuromorphiques et dans les derniers GPU de Nvidia Volta.

[91] Voir Everything you wanted to know about Deep Learning for Computer Vision but were afraid to ask (25 pages) qui décrit bien les techniques d'optimisation de la phase d'entraînement d'un réseau convolutionnel.

[92] Voir Advancing state-of-the-art image recognition with deep learning on hashtags, de Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Manohar Paluri et Laurens Van Der Maaten, mai 2018.

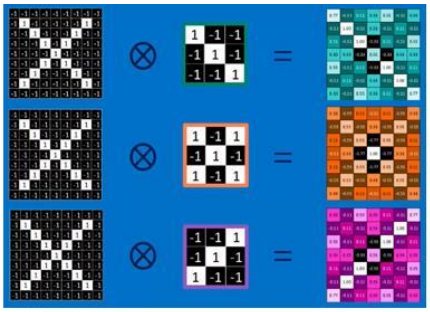
Chaque feature map générée par l'application des filtres sur l'image de départ se voit appliquée une réduction de résolution (**Pooling**) puis une suppression des valeurs négatives (**ReLU** pour Rectified Linear Units) pour réduire la quantité de travail à appliquer aux couches suivantes.

Le processus est répété sur plusieurs couches, chaque feature map issue d'un niveau devenant une image qui subit un traitement équivalent dans la couche de convolution suivante. A la fin de l'histoire, la dernière couche de feature maps est reliée à une liste de tags avec une probabilité de correspondance via quelques couches de neurones dites « **fully connected** », à savoir que tous les neurones d'une couche sont liés à ceux de la couche suivante.

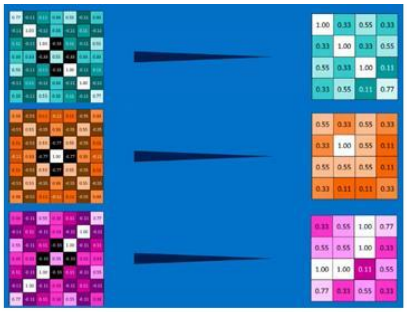
C'est là qu'un chat ou un bateau sont reconnus dans l'image. La dernière couche de cet empilement est un ensemble de neurones dont le nombre est égal au nombre de classes d'objets différents à reconnaître. Il peut être très grand mais doit rester raisonnable pour tenir compte des capacités du matériel.

Plus est grand le nombre d'objets différents que l'on veut reconnaître, plus devra être grand le nombre de paramètre du réseau de neurones. Un ConvNet de type VGG16 à cinq couches de convolution et prenant en entrée des images de 224x224 pixels en couleur (RGB) contient 138 millions de paramètres. Pour les images en couleur, les filtres de la première couche de convolution sont des matrices nxnx3, n étant la taille des filtres. La feature map résultante est une matrice carrée avec une seule couche de profondeur. Chaque point de cette matrice est le résultat de la comparaison entre le filtre à trois couleurs avec des matrices de même taille extraites de l'image d'origine. La couleur peut être représentée en mode RGB ou dans un espace colorimétrique YUV, avec une luminance, la saturation et la chrominance.

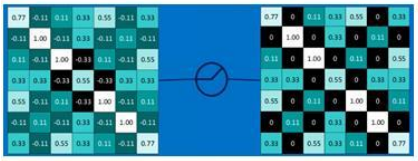
Les moteurs de reconnaissance d'images reconnaissent au grand maximum que quelques dizaines de milliers de classes d'objets dans cette dernière couche de réseaux de neurones[93]. C'est lié en particulier aux contraintes de la mémoire des GPU utilisés. Les derniers GPU Nvidia qui sont couramment utilisés pour entraîner des ConvNets disposent de 16 Go à 32 Go de mémoire.



1) feature maps : application de différents filtres à l'image de départ

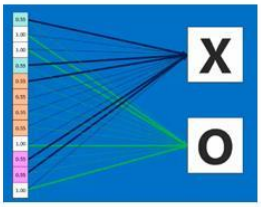


2) pooling : réduction de résolution des feature maps



3) ReLU : Rectified Linear Units, les valeurs négatives sont mises à zéro

4) empilement : de plusieurs niveaux de convolution, ReLU et pooling



5) fully connected layers : lien entre les valeurs du dernier niveau convolutionnel et le tableau des objets à reconnaitre

Voici, *ci-dessus*, un autre exemple illustré du processus des ConvNet de reconnaissance de caractères. En 1), on peut identifier la présence des diagonales et croix dans les feature maps à droite.

[93] D'ailleurs, certaines démonstrations étonnantes de reconnaissance d'objets oublient de préciser le nombre d'objets que le système peut reconnaître !

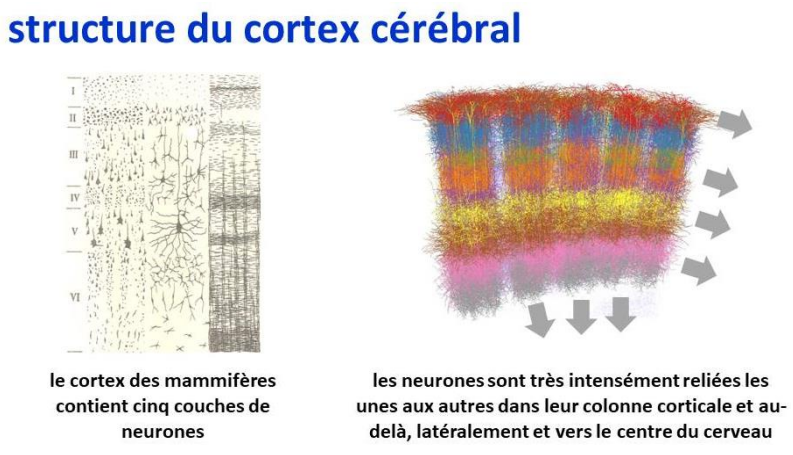
Puis le pooling en 2) pour divise par deux la résolution des feature maps, la couche ReLU qui fait un max(0, x) sur toutes les valeurs (avant ou après le pooling), puis en 5), les couches de neurones qui aboutissent au résultat final indiquant la valeur de la lettre. Selon les modèles, des variantes diverses sont introduites dans ces couches qui visent en général à augmenter le contraste de l'image traitée.

A chaque niveau d'un réseau convolutionnel, le nombre de feature maps augmente et leur taille diminue. Les feature maps étant optimisées automatiquement, leur forme n'est pas vraiment interprétable par le cerveau humain, alors que les filtres du traitement classique d'images antérieur aux réseaux convolutionnels l'étaient.

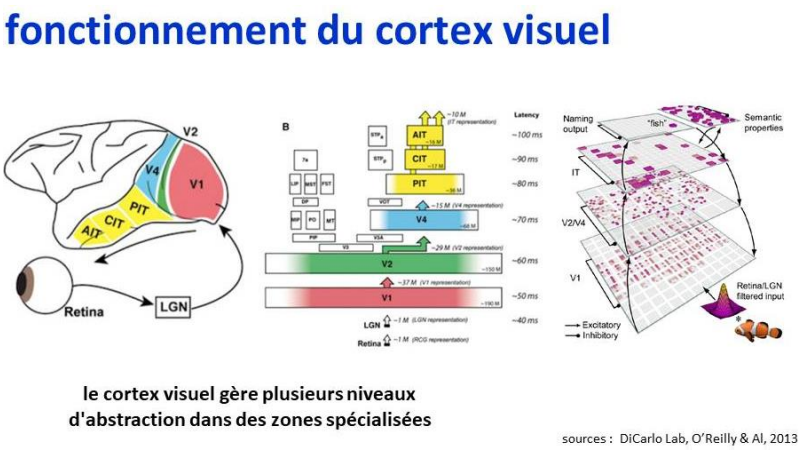
C'est la magie des ConvNets : ils créèent des niveaux de représentations hiérarchiques intermédiaires des images qui optimisent leur reconnaissance, sans que l'on puisse comprendre comment ils fonctionnent pas à pas et dans le détail[94].

D'où la fameuse « non explicabilité » des algorithmes qui inquiète nombre d'observateurs[95], ce d'autant plus qu'elle se produit aussi dans les réseaux récurrents et à mémoire qui servent principalement au traitement du langage. Mais, pour beaucoup d'applications, ce qui compte avant tout est la qualité des résultats plus que leur explicabilité. En cas de défaillance d'un réseau de neurones, l'erreur proviendra probablement d'une base d'entraînement ne couvrant pas bien l'espace des possibilités que le réseau peut rencontrer dans sa mise en production.

Nous en reparlerons plus loin au sujet du biais des données d'entraînement. Les ConvNets s'inspirent fortement du mode de fonctionnement du cortex visuel des mammifères qui est structuré, de près, dans des colonnes corticales faites de cinq couches de neurones et qui, de loin, comprend des aires spécialisées qui élèvent progressivement le niveau d'abstraction des objets reconnus (*ci-contre*) [96].



On retrouve cette architecture à cinq couches dans bon nombre de réseaux de neurones convolutionnels, sachant que ces couches peuvent elles-mêmes comprendre de nombreuses sous-couches de neurones. Par contre, contrairement au cortex humain, les ConvNets qui font de la reconnaissance d'images utilisent des représentations à très basse résolution.



[94] Par contre, comme la résolution des feature maps diminue de couche en couche, la détection de l'emplacement des macro-objets détectés est très mauvaise. Elle est même quasiment inexistante.

[95] Voir Le talon d'Achille de l'intelligence artificielle de Benoit Georges, mai 2017.

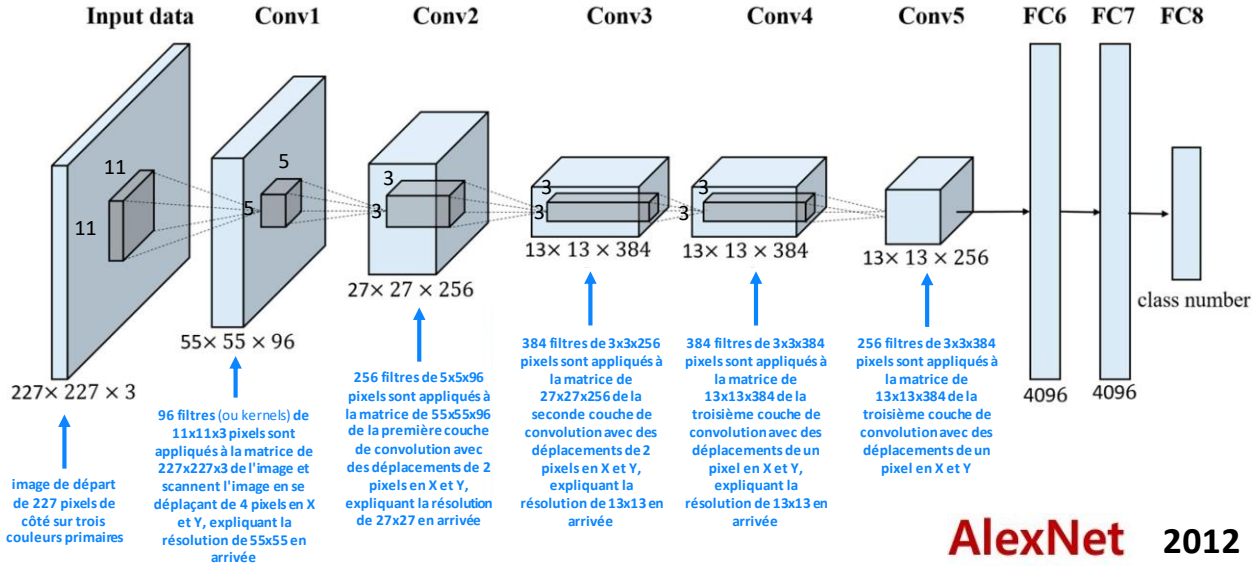
[96] Voir Receptive fields, binocular interaction and functional architecture in the cat's visual cortex de David Hubel et Torsten Wiesel, 1962.

La majorité des ConvNets se contentent d'images comprimées à une résolution voisine de 224x224 pixels et rarement au-delà de 256 pixels.

Il a fallut que je décortique le convnet **AlexaNet** qui date de 2012 pour mieux comprendre la manière dont les couches de convolution étaient reliées entre elles.

Les schémas qui décrivent ce genre de convnet peuvent facilement donner le mal de tête[97]. J'ai eu mal à la tête mais j'ai enfin compris ! En voici donc une explication.

Le principe décrit visuellement *ci-dessous*, consiste à utiliser des filtres multidimensionnels. Dans les cinq couches de convolution d'AlexNet, les filtres sont des matrices de respectivement 11x11x3 pixels pour la première couche de convolution (11 pixels de côté x 3 couleurs), 5x5x96 pixels pour la seconde (96 étant le nombre de filtres de la première couche de convolution et de feature maps qui en résultents), 3x3x256 pixels pour la troisième (256 étant le nombre de filtres et feature maps de la seconde couche de convolution), 3x3x384 pixels pour la quatrième (384 étant le nombre de filtres de la troisième couche de convolution) et 3x3x384 pixels pour la cinquième (bis repetita pour la quatrième couche de convolution).

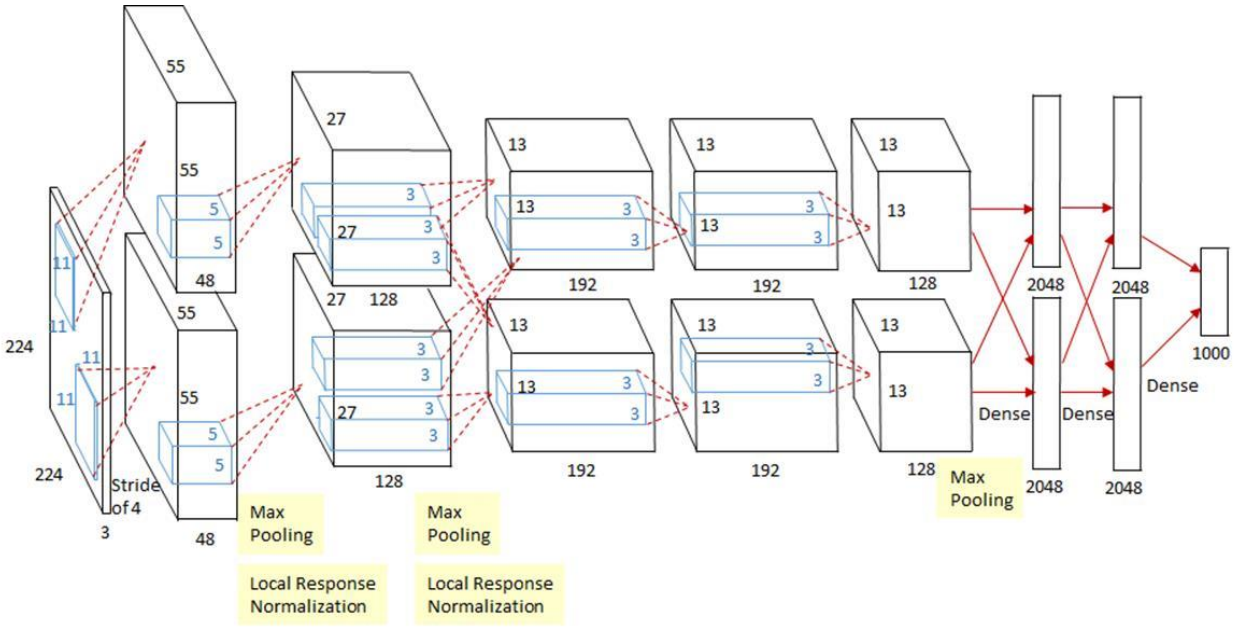


Comment passe-t-on de 227 à 55 puis à 13 pixels de côté entre l'image de départ et les couches de convolution suivantes ? Cela vient du fait que les filtres de la première convolution se déplacent de 4 pixels en 4 pixels (avec du recouvrement puisqu'ils font 11 pixels de côté), ceux de la seconde et de la troisième se déplacent de 2 pixels en 2 pixels (avec également du recouvrement puisqu'ils font 5 et 3 pixels de côté) et les deux derniers de 1 en 1 pixel. Ce déplacement est dénommé « stride » en anglais. Cette méthode est l'équivalent d'une réduction de résolution avec une couche de pooling mais elle semble plus efficace et rapide qu'un scan pixel par pixel des images dans chaque convolution puis une réduction de résolution d'un facteur 4 ou 2.

Mathématiquement, l'application d'un filtre à un morceau d'image ou de couche de convolution de la même taille consiste à multiplier les valeurs des pixels du filtre à celle de l'image ou de la couche de convolution scannée, à additionner le tout, à ajouter un éventuel coefficient de correction appelé un biais et à y appliquer une fonction d'activation qui va normaliser le résultat entre 0 et 1. C'est souvent une fonction sigmoïde. Cette opération mathématique va permettre d'identifier la dose de points communs entre le filtre et le bout de convolution analysé qui fait la même taille que le filtre. L'objet détecté est très « macro » puisque le filtre est multidimensionnel.

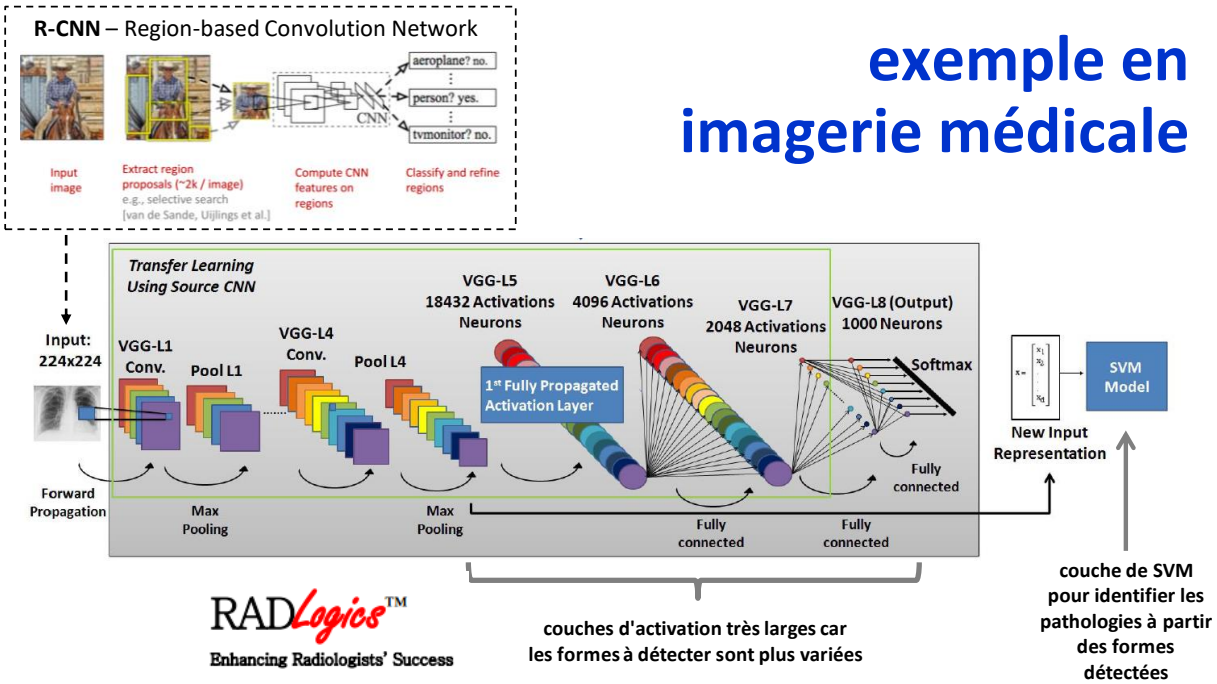
[97] Voir ImageNet Classification with Deep Convolutional Neural Networks d'Alex Krizhevsky, Ilya Sutskever et Geoffrey Hinton de l'Université de Toronto, 2012 (9 pages).

La dernière convolution d'AlexaNet génère une matrice de 13x13x256 pixels qui alimente à son tour trois couches de neurones « fully connected » permettant d'identifier 1000 objets différents.



La description d'AlexNet dans l'article d'origine est encore plus tarabiscotée, *ci-dessus*. Cela vient du fait que les convolutions sont découpées en deux parties (en haut et en bas) pour être réparties sur deux GPU. C'était l'un des premiers réseaux convolutionnels avec un entraînement distribué sur plusieurs processeurs. Aujourd'hui, on arrive à le faire sur un très grand nombre de GPUs.

L'imagerie médicale qui est plus exigeante n'échappe pas à l'exploitation d'images à basse résolution. Elle exploite généralement des réseaux de neurones de type **R-CNN** (Regional CNN) qui détectent d'abord les parties de l'image à analyser, avec une version basse résolution de l'image[98].



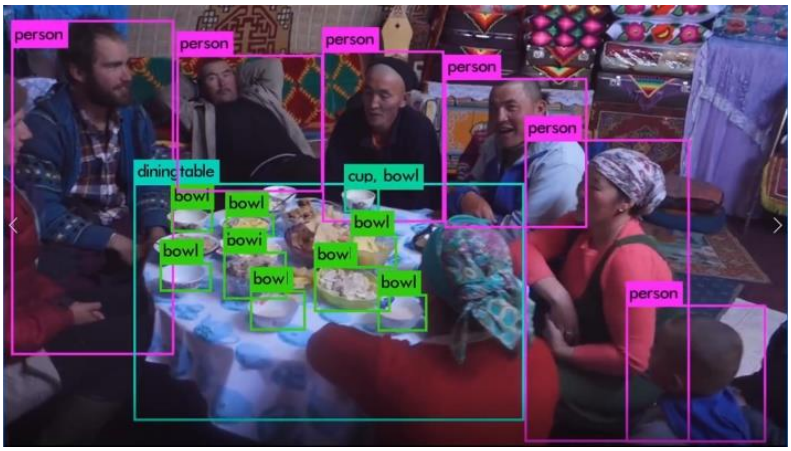
[98] Avec notamment les réseaux de neurons pyramidaux. Voir par exemple Feature Pyramid Networks for Object Detection, 2016

Il existe de nombreuses variantes de R-CNN avec les **Fast R-CNN**, **Faster R-CNN** et **Mask CNN**[99]. Chaque partie détectée est alors labellisée avec un réseau convolutionnel avec une résolution de départ de 224x224 pixels ou voisine.

En pratique, les images sont reconnues en couleur sauf lorsqu'une version monochrome de l'image est suffisante pour obtenir une reconnaissance avec un faible taux d'erreur, ce qui est souvent le cas dans l'imagerie médicale qui est fréquemment réalisée en noir et blanc (radios, IRM, échographies).

La détection d'objets multiples dans une image peut faire appel à **Yolo V3** (2016) ou « You only look once », une sorte de Convnet de détection d'objets multiples dans des images, notamment dans les vidéos.

Cela fonctionne en temps réel et jusqu'à 45 images par secondes, mais cela dépend évidemment du matériel utilisé pour les inférences.



Yolo V3 utilise 24 couches de convolution et deux couches « fully connected » en sortie avec des entrées en 224x224 pixels. La technique est voisine de celle des R-CNN[100].

Pourquoi donc 224 pixels de côté dans les CNN, R-CNN et autres Yolo ? Parce qu'on peut diviser cette valeur plusieurs fois par 2 et aboutir à 7, qui est la taille des feature maps de la dernière couche de convolution. Malgré la légère diminution de résolution entre une image et sa feature map, celle-ci reste stable. Cela vient du fait que l'on ajoute un pixel à 0 sur les bords de l'image avant de la balayer avec les filtres qui font souvent 3 pixels de côté.

La feature map de la première couche fait ainsi 224 + 2 (pixels de bord) – 3 (taille des filtres) +1 (effet de bord) = 224 pixels. Les couches suivantes font 112, 56, 28, 14 puis 7 pixels. Et voilà l'explication !

Mais il peut y en avoir des variantes avec des nombres de pixels avoisinants et des feature maps de taille variable selon les couches comme dans AlexNet (avec des filtres de 11, 5 et 3 pixels de côté) comme nous l'avons vu précédemment.

Afin de réduire la charge de calcul et de réduire la consommation d'énergie, les valeurs gérées en entrée et en sortie de neurones sont souvent encodées dans des entiers sur 4 ou 8 bits au lieu de nombres flottants 16 ou 32 bits. C'est pour cela que la performance des processeurs embarqués est souvent décrite en GigaOps et pas en GigaFlops. Les « Ops » sont des opérations sur des entiers (4, 8 ou 16 bits à préciser) et les « Flops » sur des nombres flottants. Un GigaOps n'est donc pas équivalent à un GigaFlops !

Les algorithmes utilisés sont cependant si puissants qu'ils permettent de générer des taux de reconnaissance d'images meilleurs que ceux de l'Homme ! Qu'est-ce que cela serait si la résolution utilisée était la même que dans l'œil et le cortex humains, l'œil étant doté de 90 millions de bâtonnets pour détecter la luminance et de 6,5 millions de cones pour la couleur, le tout étant connecté au cortex visuel par un nerf optique comprenant un million d'axones !

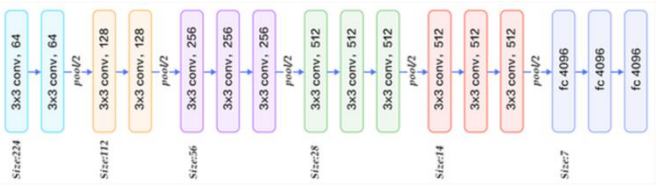
---

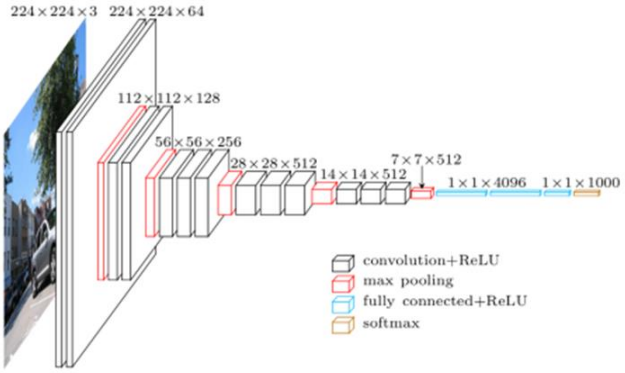
[99] Voir A Brief History of CNNs in Image Segmentation: From R-CNN to Mask R-CNN de Dhruv Parthasarathy, avril 2017.

[100] Voir You Only Look Once: Unified, Real-Time Object Detection, (10 pages) et cette vidéo. Le projet est open source.

On peut distinguer les ConvNets selon le nombre de dimensions des données reconnues : 1D (une dimension) pour le texte, la reconnaissance de genre de musique, des prévisions temporelles sur une seule variable, 2D (deux dimensions) pour les images, pour la reconnaissance de la parole qui associe fréquence audio et temps, puis 3D (trois dimensions) pour le traitement de vidéos et d'imagerie médicale 3D.

Il existe un très grand nombre d'algorithmes de ConvNets[101]. L'un des plus courants est **VGG16** (2014) qui comprend 16 couches de convolution avec des filtres de 3x3 pixels, de 64 à 512 filtres et features maps par couche et la capacité de détection de 1000 classes d'objets différentes. Le modèle est alimenté par 528 Mo de paramètres.

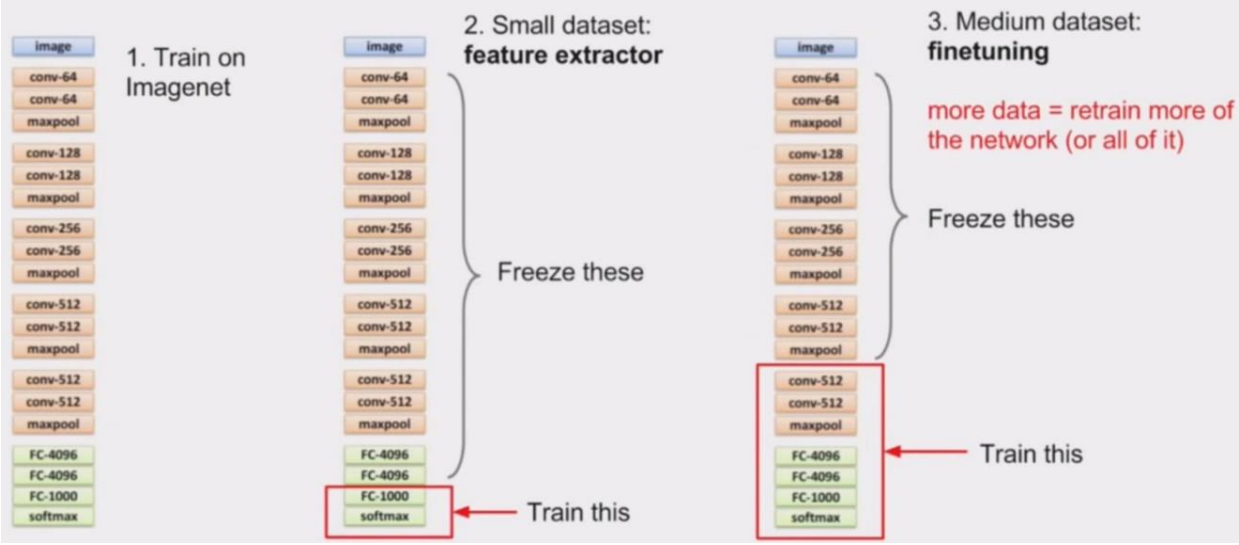




Plus on va augmenter le nombre de couches et le nombre de filtres par couches, plus on va augmenter la taille mémoire nécessaire pour la gestion de ces paramètres, de type HBM2 dans les Nvidia V100. Cela explique pourquoi les Convnets analysent des images à assez basse résolution pour le moment.

Apprentissage par transfert (1997)

L'apprentissage par transfert permet de réaliser de l'apprentissage incrémental dans les réseaux de neurones, aussi bien convolutifs et à mémoire. Il permet de transférer ce qui a été appris par un réseau dans un autre réseau, et même dans un domaine un peu différent[102]. par exemple un réseau ayant appris à reconnaître des chats sert à initialiser un réseau pour reconnaître des chiens. Cette possibilité provient de ce que les primitives de bas niveau, les filtres des premières couches, sont les mêmes pour des familles d'images du même genre.



[101] Voir ce cours de Stanford sur les CNN : CNN Architectures de Fei-Fei Li, Justin Johnson et Serena Yeung, mai 2018 (106 slides).

[102] Voir Transfer Learning de Lisa Torrey et Jude Shavlik (22 pages, non daté), Transfer Learning : réaliser de meilleures prédictions avec peu de données d'Arthur Letang, janvier 2017, A Gentle Introduction to Transfer Learning for Deep Learning de Jason Brownless, décembre 2017, Understanding, generalisation, and transfer learning in deep neural networks, février 2017, et les présentations Transfer Learning (38 slides) et Recent Advances in Transfer Learning, 2018 (28 slides).

C'est une discipline à part entière du deep learning qui a de nombreuses variantes et applications, notamment pour réduire la quantité de données nécessaires à l'apprentissage d'un réseau de neurones.

La méthode permet aussi d'entraîner un réseau de neurones de manière incrémentale, lorsque l'on ajoute par exemple des images à reconnaître dans un convnet déjà entraîné avec un jeu d'images initial. L'exemple *ci-dessous* est tiré d'une présentation d'Andrej Karpathy.

L'apprentissage par transfert (transfer learning) permet aussi de préentraîner un réseau de reconnaissance d'images avec des images générées synthétiquement dans diverses positions[103]. On l'utilise aussi pour améliorer la labellisation d'illustrations[104].

Descente stochastique de gradient (2003)

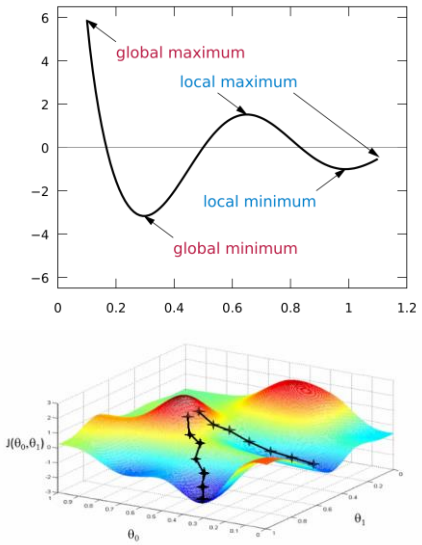
La SGD (Stochastic Gradient Descent) est une technique d'apprentissage par rétro-propagation des erreurs qui s'appuie sur l'optimisation du gradient. Pour faire simple, il s'agit d'identifier dans quelle direction faire évoluer les poids synaptiques des neurones pour atteindre leur niveau optimal dans la reconnaissance des objets en minimisant les opérations de calcul nécessaires. Le tout étant utilisé dans l'entraînement du réseau de neurones par rétropropagation d'erreurs.

Il est important de trouver le niveau optimum global, à savoir le taux d'erreur le plus bas, et pas seulement le niveau optimum local, qui est le taux d'erreur le plus bas dans les environs du poids de départ que le réseau de neurones cherche à optimiser (voir le schéma *ci-dessous* qui l'explique de manière imagée).

La technique s'applique aussi bien aux réseaux de neurones à une seule couche cachée qui font partie du domaine du marchine learning qu'aux réseaux de neurones complexes du deep learning.

Dans l'entraînement par rétropropagation d'erreurs, les poids synaptiques des neurones sont initialisés aléatoirement. On fait passer des objets d'une base de test au travers du réseau et on compare le résultat de classification en sortie avec le bon résultat dont on dispose dans la base d'entraînement (en amont, des photos et en aval, des descripteurs des objets dans les photos).

La descente de gradient évalue dans quelle direction faire évoluer les poids des synapses pour s'approcher du bon résultat. Le calcul est réalisé pour toutes les synapses et pour tous les objets du jeu d'entraînement, ce qui génère beaucoup de calculs.



la descente de gradients consiste à faire varier les poids des neurones pour trouver la valeur optimale globale, celle qui minimise le niveau d'erreur du réseau de neurones.

opération effectuée pour chaque neurone et pour chaque objet de la base de référence.

"stockastic gradient descent" est une des optimisations qui ajuste les paramètres du réseau de neurones objet par objet ou par groupe d'objet au lieu de les calculer d'un coup pour tous les objets d'entraînement à la fois

la difficulté consiste à trouver le minimum global et pas simplement un minimum local

[103] Voir par exemple Improving SAR Automatic Target Recognition Models with Transfer Learning from Simulated Data de David Malmgren-Hansen, Anders Kusk, Jørgen Dall, Allan Aasbjerg Nielsen, Rasmus Engholm et Henning Skriver, 2017 (6 pages).

[104] Voir Transfer Learning for Illustration Classification, 2018 (9 pages).

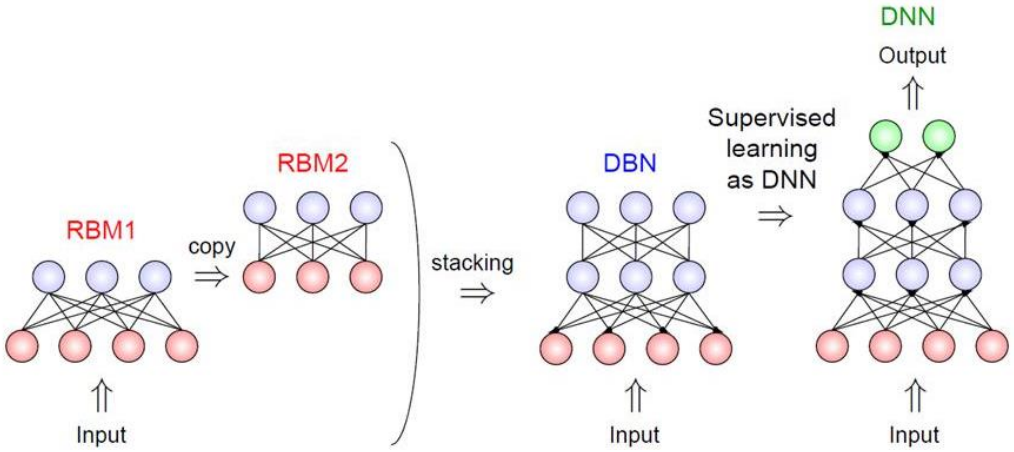
La descente stochastique de gradient est une variante de la descente de gradient qui consiste à évaluer le poids des synapses objet par objet (méthode stochastique) ou groupe d'objets (méthode mini-batch) du groupe d'objets d'entraînement au lieu de scanner entièrement la base d'entraînement (méthode dite « batch »). Cela permet de réduire la quantité de calculs à effectuer et permet de converger plus rapidement vers un réseau bien entraîné.

Cette technique d'entraînement est très efficace pour générer un réseau de neurones capable de générer des résultats avec un faible taux d'erreurs. Elle est cependant très consommatrice de ressources machines et de temps. D'où l'intérêt de l'optimiser et de s'appuyer sur des ressources matérielles de plus en plus puissantes, comme les ordinateurs à base de GPU ou de processeurs neuromorphiques que nous étudions plus loin dans ce document.

Deep beliefs networks (2006)

Les DBN sont issus des travaux des canadiens Geoffrey Hinton et Simon Osindero et du Singapourien Yee-Whye Teh publiés dans **A Fast Learning Algorithm For Deep Belief Nets**. Ils optimisent le fonctionnement des réseaux neuronaux multicouches en gérant leur apprentissage couche par couche, indépendamment les unes des autres.

Ce sont en quelque sorte des machines de Boltzmann restreintes empilées les unes sur les autres, étape par étape pour ce qui est de l'entraînement.



Le concept général du deep learning a été ensuite formalisé par ce même Geoffrey Hinton en 2007 dans **Learning multiple layers of representation**.

Notons que Geoff Hinton s'appuyait en partie sur les travaux du français **Yann LeCun** (en 1989) qui dirige maintenant le laboratoire de recherche en IA de Facebook et de l'allemand **Jürgen Schmidhuber** (1992, créateur des LSTM) dont deux des anciens étudiants ont créé la start-up **DeepMind**, maintenant filiale de Google. Sachant que Yann LeCun était lui-même un ancien post-doctorant dans le laboratoire de Geoff Hinton. Un bien petit monde !

Geoffrey Hinton[105] travaille pour **Google** depuis 2013, pas loin du légendaire **Jeff Dean**[106], arrivé en 1999 et qui planche maintenant aussi sur le deep learning**.** On peut aussi citer le français **Stéphane Mallat** qui a aussi contribué au développement des réseaux convolutionnels et à l'explication de leurs sous-jacents mathématiques[107].

---

[105] Voir Is AI Riding a One-Trick Pony? de James Somers, septembre 2017, MIT Technology Review, qui montre à quel point Geoff Hinton est central dans l'histoire récente de l'IA.

[106] Co-créateur entre autres de choses de deux outils clés des traitements distribués MapReduce et BigTable, ainsi que du crawler de Google Search et d'AdSense.

[107] Sa conférence délivrée dans la Chaire du Collège de France de Yann LeCun fournit des éclaircissements sur le fonctionnement des réseaux convolutionnels. Mais il faut s'accrocher pour suivre ! Voir les Mystères mathématiques des réseaux de neurones convolutionnels, 19 février 2016.

Autoencodeurs empilés (2007)

Les *stacked autoencoders* sont couramment associés aux deep belief networks et aux réseaux convolutionnels. Il s'agit d'utiliser des couches cachées de neurones qui encodent les données dans un espace dimentionnel réduit et des couches de neurones qui permettent ensuite de reconstituer les variables en entrées, en sortie de cette couche.

Cette technique est utilisée dans l'apprentissage non supervisé des réseaux de neurones pour identifier des variables ou fonctions cachées. Elle peut notamment servir à débruiter des images.

Cette technique est donc une technique de réduction de dimensions. La méthode mathématique employée peut être la PCA (Principal Components Analysis) que nous avons rapidement vue dans la partie sur le machine learning ou la Singular Value Decomposition (SVD), autre méthode voisine de la PCA. En fait l'autoencodeur réalise une sorte de PCA non linéaire[108].

Il existe bien entendu diverses variantes d'autoencodeurs comme les DAE (denoising autoencoders), les SAE (sparse autoencoders) et les VE (variational autoencoders).

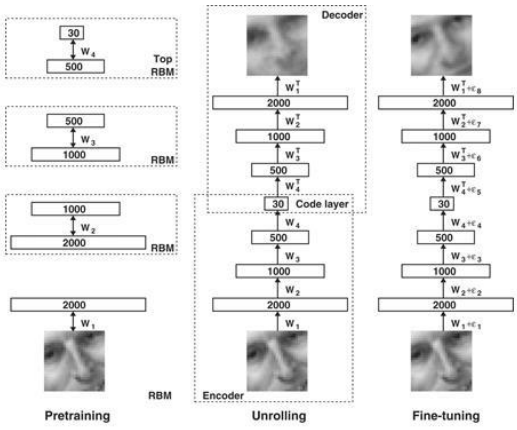
les autoencodeurs empilés sont des couches cachées de neurones qui encodent les données dans un espace dimentionnel réduit

l'autoencodeur est capable de reconstituer les variables en entrées en sortie de cette couche

utilisé en mode d'apprentissage non supervisé pour trouver des variables cachées, y compris dans les CNN, la couche de réencodage n'étant pas utilisée en production

peut notamment servir à débruiter des images (de caractères, ...) dans des autoencodeurs empilés

la méthode employée peut-être la PCA (Principal Components Analysis)



Modèles génératifs (2014)

Les réseaux de neurones génératifs sont des réseaux de neurones convolutionnels ou récurrents générant du contenu à partir de contenus existants.

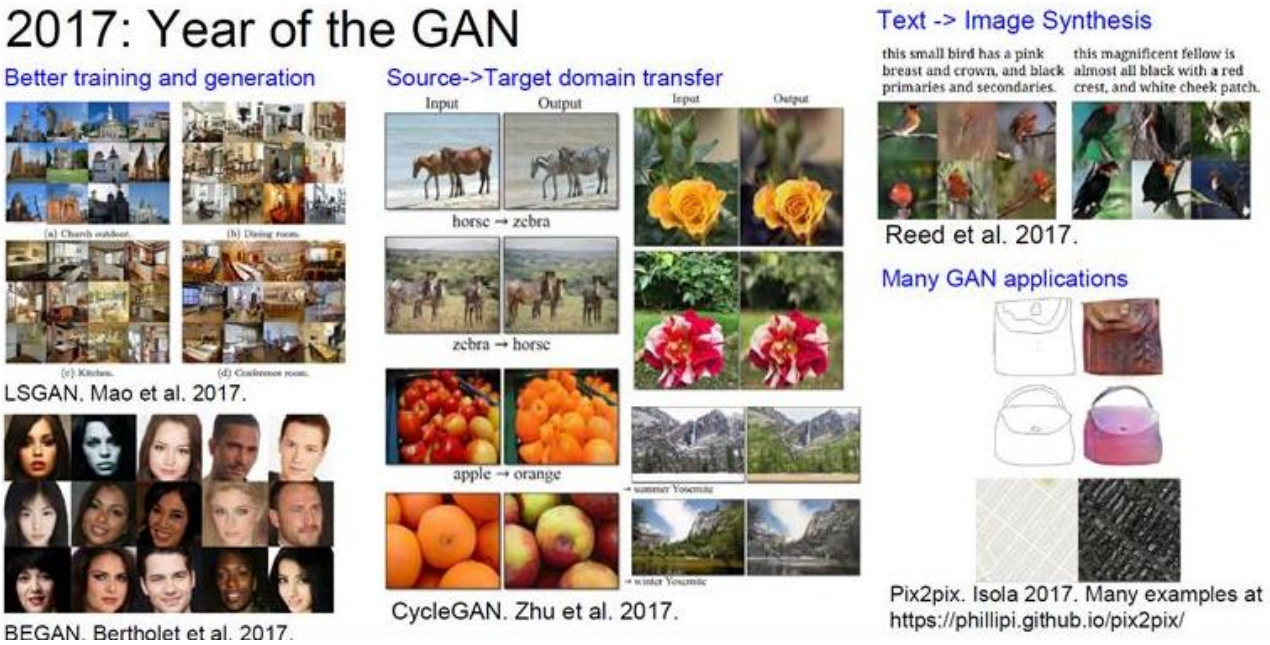
Ils font des prévisions, d'images vidéo suivantes d'une vidéo donnée, ils colorient des images en noir et blanc[109] et ils remplissent des images dont il manque des morceaux. Ils peuvent aussi servir à améliorer les dialogues de chatbots.

Les principales techniques utilisées sont les **GAN** (Generative Adversarial Networks), apparus en 2014 et perfectionnés en particulier en 2016. Ces sont des réseaux de neurones non supervisés capables de générer des contenus en s'appuyant sur des générateurs à base de réseaux de convolution inversés, sortes de stacked autoencoders multiples et de discriminateurs qui permettent d'identifier les contenus générés les plus plausibles, qui sont reconnus par un réseau de neurones de labellisation d'objets.

---

[108] Voir Auto-association by multilayer perceptrons and singular value decomposition, de Hervé Bourlard et Y. Kamp, 1988.

[109] Les exemples du slide ci-dessous viennent de : Generative Models de Fei-Fei Li & Justin Johnson & Serena Yeung, 2017.



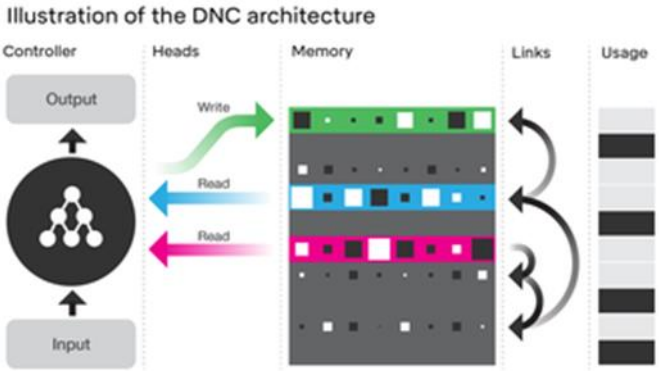
Les variantes de réseaux génératifs sont de plus en plus nombreuses avec notamment les **VAE** (Variational Autoencoders, 2014), les **Adversarial Eutoencoders** (2015), les **InfoGAN** (2016), les **CycleGAN** (2017), **Wasserstein GAN** (2017), **PixelRNN / PixelCNN** (2016) et **GMMN** (Generative moment matching networks, 2015).

Je traite des GANs dans la rubrique sur les modèles génératifs dans la partie liée à la vision car c'est le principal domaine d'application des GANs[110].

Differential Neural Computers (2016)

Les DNC sont des réseaux de neurones récurrents utilisant une mémoire autoassociative, créés par Alex Graves et son équipe de DeepMind[111]. Ils gèrent les relations entre composantes d'une mémoire long terme.

A ce jour, ils ne servent pas à grand-chose mais pourraient servir à intégrer du raisonnement symbolique dans les réseaux de neurones. Les DNC peuvent répondre à des questions sur des données structurées complexes et structurées en arbres et même résoudre certains types de puzzles. On associe ces DNC à la programmation différentielle dont l'usage est actuellement en plein essor[112]. Ils imitent aussi le fonctionnement de l'hypocampe du cerveau humain.



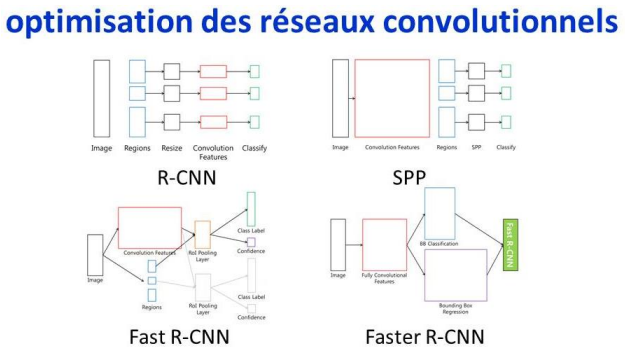
[110] Voir aussi An applied introduction to generative adversarial networks, décembre 2017.

[111] Voir Hybrid computing using a neural network with dynamic external memory, 2016 (23 pages), Differentiable Neural Computers d'Alex Graves, 2016 (16 slides) et Differentiable neural computers, 2016, du même auteur, et What are differentiable neural computers de Heidelberg.ai, 2017 (62 slides). Sinon, Hardware Accelerator for Differentiable Neural Computer and Its FPGA Implementation, 2017 (7 pages) décrit la manière de transposer les DNC dans du hardware.

[112] Voir notamment Diffential programming de Atilim Güneş Baydin, 2016 (69 slides).

Autres méthodes de deep learning *(2015-2017)*

S'en suivirent plus récemment de nombreuses variantes de réseaux de neurones, surtout à base de réseaux convolutionnels (*ci-dessous à droite*), destinées à optimiser les performances, en particulier, celles de l'entraînement des réseaux. En effet, c'est la partie qui est la plus consommatrice de ressources machines dans les réseaux de neurones. Une fois un réseau entraîné, il exécute ses prévisions bien plus rapidement.

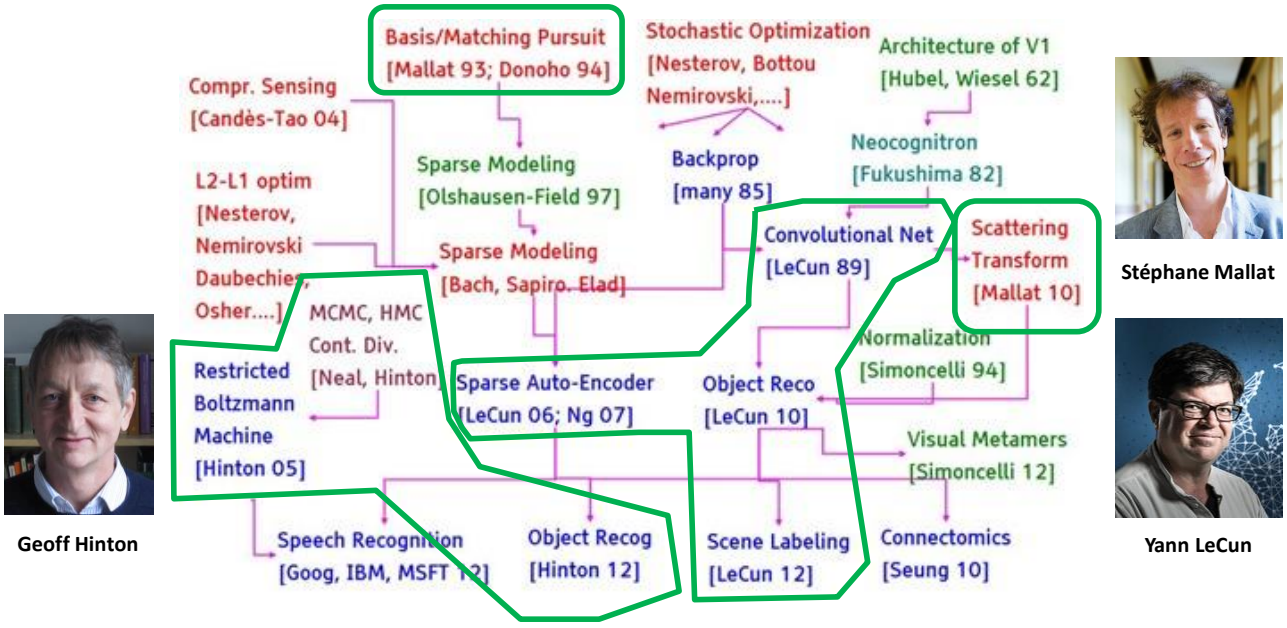


On voit aussi émerger des réseaux de **deep learning évolutifs** dont l'architecture peut évoluer de manière itérative[113].

Le schéma ci-dessous illustre cette longue chaine de progrès, qui ne s'est d'ailleurs pas arrêtée en 2012 et poursuit encore son chemin aujourd'hui. Elle s'est même accélérée avec l'arrivée dans le domaine de très nombreux chercheurs chinois. 2012 est une date intéressante, celle de la création du réseau de neurones convolutionnel **AlexNet** par l'équipe de Geoff Hinton de l'Université de Toronto. Il était entraîné sur 1,3 millions d'images de la base de référence **ImageNet** avec 1000 classes d'images différentes et générait des taux d'erreurs bas pour l'époque (18,9% pour le top 5).

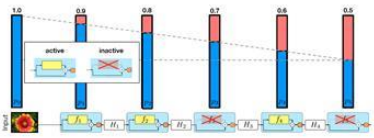
Le réseau de neurones exploitait 60 millions de paramètres et 500 000 neurones répartis dans cinq couches de convolution. Le tout exploitait des cartes GPU Nvidia pour accélérer les traitements, surtout pendant la phase d'entraînement qui durait des semaines[114].

Le deep learning et les réseaux de neurones convolutionnels ont continué de progresser pas à pas, avec la création d'innombrables variantes conceptuelles pour améliorer leurs capacités d'apprentissage, de mémorisation et leur performance d'entraînement. L'IA progresse d'ailleurs régulièrement et de manière plutôt décentralisée, avec des dizaines de chercheurs contribuant à faire avancer l'état de l'art. Le tout, pendant que les GPU et processeurs neuromorphiques progressent tout autant de leur côté.

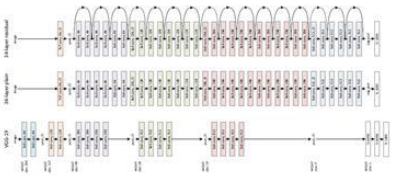


[113] Voir Neuroevolution: A different kind of deep learning, juillet 2017.

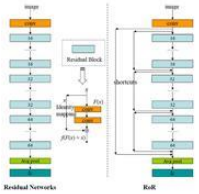
[114] Voir ImageNet Classification with Deep Convolutional Neural Networks, 2012.



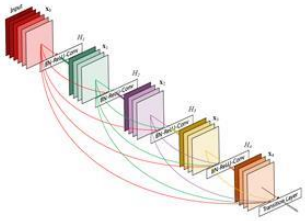
Stochastic Residual Net, 2016
accélère entrainement



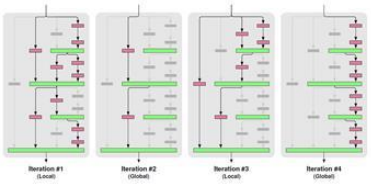
Residual Net, 2016
duplication de couches, améliore entrainement



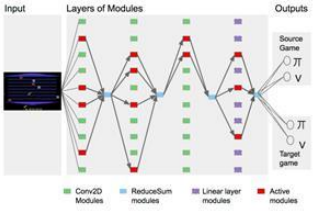
RoR, 2016
variante de ResNet



DenseNet, 2016
évite perte de gradient



FractalNel, 2016
densification de réseau



PathNet, 2016
réseau de réseau

Nous avons ainsi par exemple vu apparaître quelques avancées conceptuelles clés depuis 2015 :

- **ResNet**[115] (2015), un réseau de neurones profonds, c'est-à-dire avec de nombreuses couches, 152 !, mais qui en optimise l'entraînement. Il sert à réduire la perte de gradients dans le processus d'entraînement. Il permettait d'atteindre un taux d'erreur de 5,7% sur la base de tests ImageNet.
- **Stockastic Residual Net**[116] (2016), qui optimise les réseaux de neurones en court-circuitant certaines couches pendant l'entraînement pour le rendre plus rapide.
- **FractalNet**[117] (2016), qui utilise le concept des fractales pour densifier un réseau de neurones convolutionnel en répliquant certaines couches et en utilisant plusieurs circuits différents pour l'optimisation de chaque convolution.
- **DenseNet**[118] (2016), une variante des ConvNets où chaque feature map est injectée en entrée de toutes les couches convolutionnelles suivantes et pas seulement de la suivante, évitant le syndrome de la perte de gradient qui affecte les ConvNets lors de leur entraînement.
- **PathNet**[119] (2016), un réseau de neurones, chaque neurone étant un réseau convolutionnel, dont l'usage est optimisé automatiquement. C'est une création de DeepMind.

[115] Voir Deep Residual Learning for Image Recognition, de Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun, 2015. ResNet a été développé par une équipe de Microsoft Research en Chine. ResNet serait utilisé dans AlphaGo Zero, développé par DeepMind en 2017.

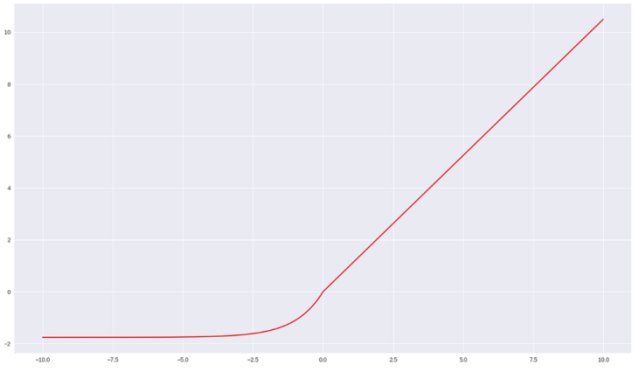
[116] Voir Deep Networks with Stochastic Depth de Gao Huang, Yu Sun, Zhuang Liuy, Daniel Sedra, et Kilian Q. Weinberger, 2016.

[117] Voir FractalNet : Ultra-deep neural networks without residuals, de Gustav Larsson, Michael Maire et Gregory Shakhnarovitch, 2016.

[118] Voir Densely Connected Convolutional Networks de Gao Huang, Zhuang Liu et Laurens van der Maaten, 2016, révisé en 2017.

[119] Voir PathNet: Evolution Channels Gradient Descent in Super Neural Networks, de Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel et Daan Wierstra, janvier 2017.

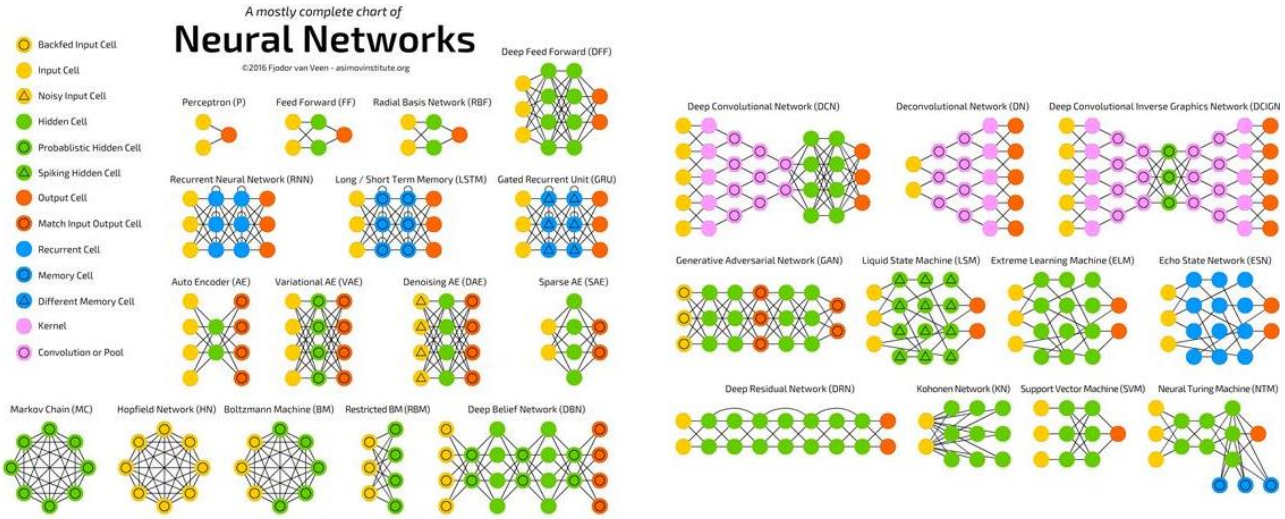
- **SNN** (2017) sont les Self-Normalizing Neural Networks[120], des variantes de ConvNets qui utilisent une function d'activation particulière, le **SELU** (Scaled Exponential Linear Units) qui intègre une function de normalisation qui n'est pas très éloignée d'une RELU et d'une ELU en étant légèrement négative pour les valeurs négatives au lieu d'être à zéro. Qui plus est, les valeurs des poids sont initialisées à 0 a lieu de 0,5. Pourquoi donc ? Pour permettre une convergence plus rapide du réseau de neurones lors de l'entraînement par rétropropagation.



Une fonction d'activation SELU permet d'améliorer d'un facteur 10 le temps d'entraînement de certains types de réseaux de neurones convolutionnels.

- **Capsule Networks** (2017), déjà cités avec plus de détails au-dessus. Cependant, je n'en ai pas découvert d'usages pratiques depuis leur apparition en décembre 2017.
- **Mixture of Expert Layer**[121] (2017), un nouveau modèle de réseau de neurones multicouches créé par une équipe de Google Brain pilotée par Geoff Hinton. C'est un réseau neuronal géant dont chaque neurone est en fait un sous-réseau neuronal. Le modèle, différent de Pathnet, sert surtout à améliorer les outils de traitement du langage comme la traduction.
- **Tree-CNN** (2018) est une des methodes les plus récentes permettant l'entraînement incrémental d'un Convnet[122]. En effet, comme un Convnet demande d'énormes ressources machines pour être entraîné, des méthodes spécifiques sont nécessaires pour modifier cet entraînement lorsque, par exemple, des images sont ajoutées voire même supprimées d'un jeu de test.

A chaque fois, ces différents réseaux ont été entraînés avec les mêmes sources de données comme la base **ImageNet**, pour reconnaître avec le taux d'erreurs le plus faible les images de test et aussi en économisant au mieux les ressources machine. L'autre point commun de ces avancées récentes est qu'elles proviennent souvent de chercheurs et étudiants chinois… installés surtout aux USA.



[120] Voir Self-Normalizing Neural Networks de Günter Klambauer, Thomas Unterthiner et Andreas Mayr (Autriche), 2017 (102 pages).

[121] Voir Outrageously large neural networks : the sparsely-gated mixture-of-experts layer de Geoffrey Hinton, Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le et Jeff Dean, janvier 2017.

[122] Voir A Hierarchical Deep Convolutional Neural Network for Incremental Learning, mai 2018 (12 pages).

Aujourd'hui, les taux d'erreurs sont inférieurs à ceux de l'Homme ce qui explique pourquoi il est souvent dit qu'en matière d'imagerie médicale, les médecins spécialistes sont dépassés. Même si ce propos est exagéré car la reconnaissance d'images ne correspond qu'à une partie seulement de leur expertise.

A ceci près que ces réseaux de neurones ont été entraînés avec des bases d'images taggées reflétant le savoir existant des spécialistes. La connaissance de l'IA ne tombe pas du ciel !

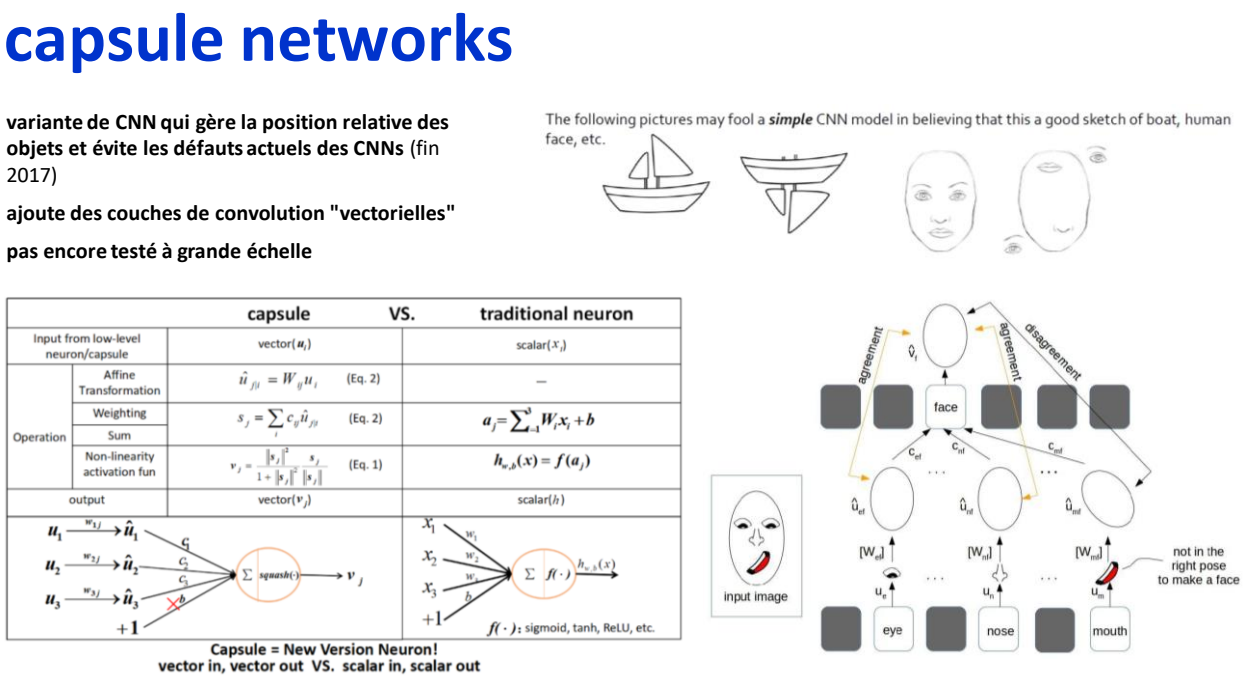
La cartographie *ci-dessus* du « zoo » des réseaux de neurones[123] illustre bien leur diversité sachant que leur assemblage peut donner ensuite lieu à beaucoup de créativité en fonction des besoins.

Capsule Network (2017)

Les Capsule Networks on été présentés fin 2017 par Geoff Hinton[124], le père du deep learning comme un moyen de permettre aux réseaux convolutifs de ne pas perdre la spatialité relative des objets intermédiaires qu'ils découvrent.

Les Capsule Networks complètent les convnets avec des couches vectorielles gérant la position relative des sous-objets découverts dans chaque couche de convolution. Pour l'instant, les Capsule Networks n'ont été testés que pour reconnaître des lettres et ils présentent l'inconvénient de nécessiter un très grand nombre de paramètres rendant leur entraînement peut-être difficile à réaliser sur les processeurs et mémoires d'aujourd'hui.

Des déclinaisons des Capsule Networks ont depuis été proposées pour améliorer la classification d'objets[125], celle de tumeurs cancéreuses du cerveau[126] et pour l'analyse de sentiments[127].



[123] Cette cartographie provient de The Neural Network Zoo de l'Institut Asimov, septembre 2016. Chaque type de réseau de neurone est bien décrit un par un. Malheureusement, cette cartographie ne semble pas avoir été mise à jour depuis 2016.

[124] Voir Dynamic Routing Between Capsules de Sara Sabour, Nicholas Frosst et Geoffrey Hinton, novembre 2017 (11 pages), cette vidéo d'Aurélien Géron qui en décrit bien le principe, 'Godfather' of deep learning is reimagining AI de Chris Sorensen, novembre 2017 et cette vulgarisation technique de haut vol Understanding Hinton's Capsule Networks en quatre partie de Max Pechyonkin.

[125] Voir Capsules for Object Segmentation de Rodney LaLonde et Ulas Bagci, avril 2018 (9 pages).

[126] Voir Brain tumor type classification via capsule networks de Parnian Afshar, Arash Mohammadi et Konstantinos Plataniotis, mars 2018 (5 pages).

[127] Voir Sentiment Analysis by Capsules, avril 2018 (10 pages).

Modes d'apprentissage

Comme pour le machine learning, l'apprentissage de solutions de deep learning suit l'une des approches suivantes :

- L'**apprentissage supervisé** qui repose sur l'entraînement d'un réseau avec un jeu de données d'entraînement qui est associé à une donnée de résultat souhaité. Pour la reconnaissance d'images, il s'agit des descriptifs d'objets contenus par les images ou leur classe. Pour de la traduction automatique, ce sont des couples de phrases traduites d'une langue à l'autre. La labellisation est généralement manuelle et d'origine humaine. Tout du moins, dès lors que l'on exploite du langage. Le tagging de données peut exploiter des sources non humaines, comme des capteurs sensoriels divers.
- L'**apprentissage non supervisé** qui est utilisé dans certains types de réseaux de neurones de deep learning ou certaines parties de réseaux, comme les stacked autoencoders qui permettent d'identifier automatiquement des patterns dans des objets et de réaliser du clustering automatique d'objets. C'est un moyen de découvrir des classes d'objets alors que dans l'apprentissage supervisé, on entraîne le modèle avec des classes préétablies manuellement. Cet apprentissage non supervisé ne va pas identifier automatiquement le nom des classes identifiées. L'apprentissage totalement non supervisé est plus que rare. Le non supervisé est souvent le complément du supervisé.

  L'apprentissage non supervisé semble indiquer qu'une machine peut faire preuve de créativité ou d'intuition. En fait, il n'en est rien. L'apprentissage non supervisé permet d'identifier automatiquement des variables discriminantes d'un élément spécifique. Dans les chatbots et le traitement du langage, l'apprentissage non supervisé est une forme d'apprentissage par renforcement. C'est le cas si le chatbot s'entraîne à mieux répondre en examinant la manière dont évoluent les conversations avec les utilisateurs en fonction de leurs réponses. Cela implique donc encore une boucle avec des humains.
- L'**apprentissage par renforcement** qui consiste à faire évoluer un modèle en fonction de retours externes, en général avec le monde physique. Cette méthode vise à optimiser une récompense quantitative (reward) obtenue par le système en fonction de ses interactions avec son environnement. C'est une technique qui est par exemple utilisée pour optimiser le réalisme des dialogues de chatbots, dans les jeux vidéo ou en robotique.

  Elle l'est dans les robots qui apprennent à éviter les obstacles ou à réaliser des tâches mécaniques en tâtonnant. L'agent à entraîner par renforcement cherche à maximiser par itérations successives une récompense qui est incarnée par sa performance, telle que le temps pour réaliser une tâche donnée ou la qualité de cette tâche. L'apprentissage par renforcement peut nécessiter un grand nombre d'essais et de données[128].

Applications du deep learning

Depuis 2012, le deep learning est mis à toutes les sauces, la plus symbolique étant la victoire de **DeepMind** contre le champion du monde de Go à la mi-mars 2016.

Le deep learning est surtout utilisé aujourd'hui pour la reconnaissance des formes dans les images et celle de la parole, donc dans les sens artificiels.

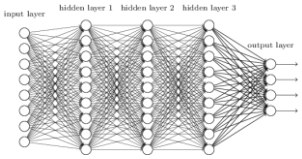
Il peut aussi servir à exploiter des données textuelles non structurées et à alimenter des bases de connaissances qui elles-mêmes seront exploitées par des moteurs de règles dans des systèmes experts utilisant une logique formelle ! IBM liste quelques-unes de ces applications dans **son marketing**.

[128] Voir ce cours en 10 sessions de 1h30 en anglais sur l'apprentissage par renforcement de David Silver.

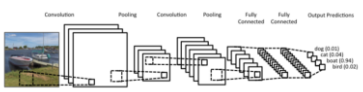
On y retrouve des études de cas dans l'éducation pour créer des MOOC auto-adaptatifs, dans le retail avec un assistant d'achats, dans la santé avec la personnalisation de traitements contre certains cancers ou encore dans l'analyse de diverses données dans la smart city.

Pour comprendre le fonctionnement du deep learning dans le détail, il faut avoir beaucoup de temps et un bon bagage mathématique et logique ! On peut commencer par parcourir **Deep Learning in Neural Networks** de Jürgen Schmidhuber, publié en 2014 qui fait 88 pages dont 53 de bibliographie ou bien **Neural Networks and Deep Learning**, un livre gratuit en ligne qui expose les principes du deep learning. Il explique notamment pourquoi l'auto-apprentissage est difficile.

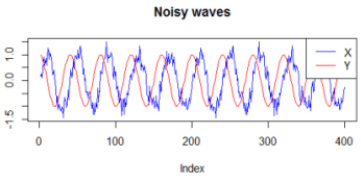
Il y a aussi **Deep Learning Methods and Applications** publié par Microsoft Research (197 pages) qui démarre en vulgarisant assez bien le sujet. Mais elle commence à dater, la troisième et dernière édition étant de 2009. Il y a aussi la masse **Deep Learning** de Ian Goodfellow, Yoshua Bengio et Aaron Courville, de 802 pages[129].



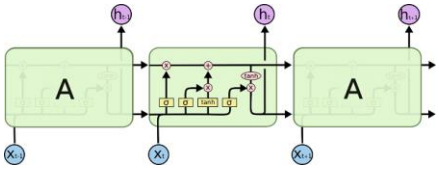
fully connected
classification
et prédictions



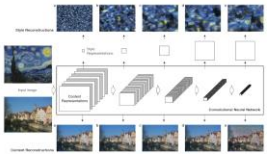
convolutionnels
spatial
reconnaissance images



récurrents
temporels
ECG, finance, bruit



LSTM
contexte - bidirectionnel
traduction, dialogue, recherche



transfer networks
apprentissage incrémental
changement de domaine

génératifs
variations – augmentation
modification d'images et de textes

Vous pouvez aussi visionner la **conférence inaugurale** de Yann LeCun au Collège de France en février 2016 où il excelle dans la vulgarisation même si l'on peut avoir du mal à suivre jusqu'à la fin la première fois.

Le deep learning est très coûteux en ressources marchines, surtout pendant les phases d'entraînement. Nous avons vu que celui-ci passe par l'optimisation des poids de centaines de millions de neurones qui doit être réalisée en testant chaque objet de référence en entrée et il peut y en avoir plusieurs millions. Chaque traitement d'une image de référence peut nécessiter des milliards d'opérations. Les processeurs traditionnels ne sont pas bien adaptés à ces traitements. En effet, ils vont tester et adapter séquentiellement le poids des synapses de chaque neurone et la construction des « feature maps » des couches convolutionnelles.

Du côté du livre des records :

- En 2011, **Google Deep Brain** reconnaissait des chats dans des vidéos YouTube avec un réseau comprenant 1,7 milliards de connexions, 10 millions d'images de 200x200 pixels, 1000 machines et 16 000 cœurs, 170 serveurs, 20 000 catégories d'objets et 3 jours de calcul[130].

[129] Téléchargeable ici : http://www.iro.umontreal.ca/~bengioy/talks/lisbon-mlss-19juillet2015.pdf.

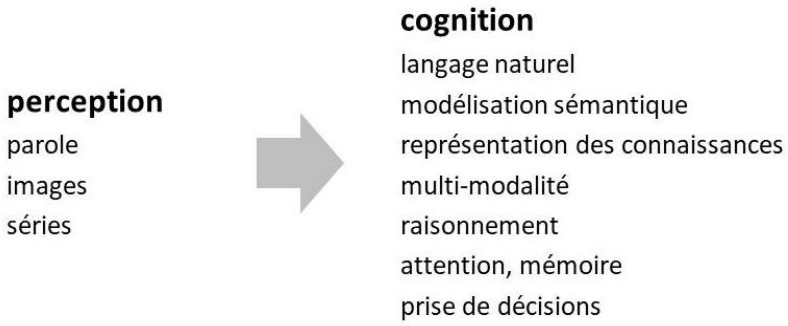
[130] Voir Google's artificial brain learns to find cat videos, Wired, 2012.

- En 2013, une équipe de **Stanford** sous la direction d'Andrew Ng créait un réseau de neurones de reconnaissance d'images de 11,2 milliards de paramètres tournant sur 16 serveurs à base de GPU Nvidia[131].
- En 2015, le **Lawrence Livermore Lab** créait un système gérant 15 milliards de paramètres[132] pour reconnaître des objets dans une base de 100 millions d'images issue de Flickr.
- Toujours en 2015, la startup **Digital Reasoning** de Nashville exploitait un réseau de neurones de traitement du langage cherchant des analogies parmi 20 000 mots et avec 160 milliards de paramètres, entraîné sur trois serveurs en une journée[133]. Avec un taux d'erreur de moins de 15%, un record à l'époque.
- Encore en 2015, on passait à la reconnaissance de visages avec **Nvidia**, toujours sur 100 millions d'images, avec 10 niveaux de neurones, un milliard de paramètres, 30 exaflops et 30 GPU-jours de calculs pour l'entraînement[134].
- En 2017, ce sont les **réseaux génératifs** qui ont le plus impressionné avec leurs capacités à générer des visages de synthèse, à compléter des images incomplètes ou à coloriser des photos noir et blanc.
- En 2018, **Samsung** présentait une TV 8K censée faire de l'upscaling d'images SD, HD et UHD en 8K grâce à des réseaux de neurones génératifs. On est en droit de douter mais avec une bonne base d'entraînement, pourquoi pas.

Ces performances vertigineuses s'expliquent notamment par la vague de l'usage de GPU et de processeurs neuromorphiques dont la structure interne est plus adaptée aux calculs des réseaux de neurones que les CPU traditionnels. Ces processeurs savent paralléliser les calculs et multiplier des matrices entre elles, ce qui est utile pour les réseaux de neurones convolutionnels. Nous verrons dans une partie suivante comment progressent ces GPU, surtout issus de Nvidia, et les processeurs neuromorphiques.

Jusqu'à présent, nous avons évoqué les applications du deep learning dans la reconnaissance des formes. Le deep learning a-t-il d'autres usages, notamment dans le cognitif et dans l'intelligence symbolique, jusqu'ici l'apanage des systèmes experts ? Oui, dans une certaine mesure.

Ces techniques dites cognitives sont des techniques avancées de traitement du langage avec une vision plus statistique que logique. Il reste un sacré chemin à parcourir pour intégrer de la logique et du raisonnement dans les réseaux de neurones. Nous en parlerons dans la partie dédiée à l'intelligence artificielle générale.



De même, l'application du deep learning pour des besoins industriels comme la maintenance prédictive, le marketing, l'analyse des bases clients est tout à fait embryonnaire.

---

[131] Voir GPU-Accelerated Machine Learning and Data Mining Poised to Dramatically Improve Object, Speech, Audio, Image and Video Recognition Capabilities, Nvidia, 2013.

[132] Voir Large-scaled deep learning ont the YFCC100M dataset, 2015.

[133] Voir Biggest Neural Network Ever Pushes AI Deep Learning, et Modeling Order in Neural Word Embeddings at Scale, 2015.

[134] Voir Deep learning image classification, Nvidia, 2016.

Outils du deep learning

Poursuivons cette partie sur le deep learning en évoquant l'offre des outils de création des solutions les mettant en œuvre. Il s'agit d'outils de développement exploitant des langages déclaratifs comme Python. Ils permettent de créer des modèles de réseaux de neurones avec leurs différentes couches.

La programmation consiste surtout à définir la structure du réseau de neurones : le nombre de couches cachées, la taille des filtres et des feature maps pour les réseaux de neurones convolutionnels, les fonctions de pooling (réduction de résolution), puis à déclencher son entraînement avec une boucle de programme qui va scanner un jeu d'entraînement taggé et faire de la rétropropagation de gradient dans le réseau de neurones. Une fois entraîné, on évalue le taux d'erreurs généré sur un jeu de test et on affine tous les éléments ci-dessus dans ce que l'on appelle l'optimisation des hyperparamètres.

Les outils disponibles pour créer des solutions de deep learning sont le plus souvent disponibles en open source, installables sur les machines et serveurs des utilisateurs ou accessibles via des ressources serveur en cloud.

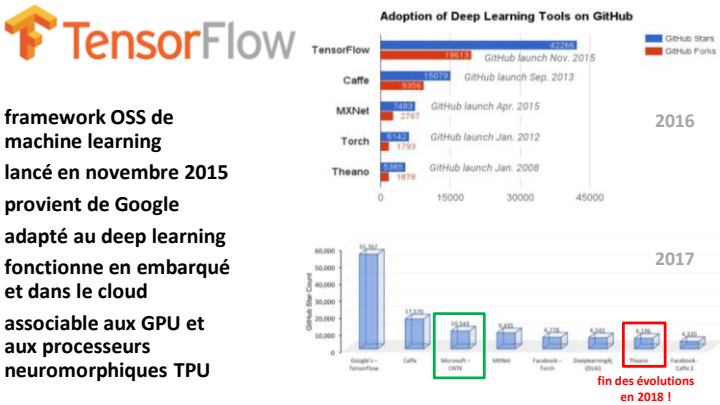
Les grands acteurs du numérique proposent tous leurs frameworks et outils open source de création de réseaux de neurones : TensorFlow chez Google, Torch chez Facebook, Cortana NTK chez Microsoft, la plateforme Watson chez IBM ou encore DSSTNE chez Amazon. Mais les startups ne sont pas en reste, comme H2O, même si les outils de développement open source semblent avoir largement pris le dessus comme c'est le cas pour le développement Internet depuis 1995.

Les modèles de réseaux de neurones se définissent soit avec des fichiers de configuration (Caffe, CNTK) soit par langage de programmation et notamment Python (Torch, Theano, TensorFlow) ou encore Lea (pour Torch). Python est le langage le plus utilisé dans ce domaine.

Ça tombe bien car il sert aussi à développer la partie back-end de nombreux sites web. Il est d'ailleurs aussi utilisé de manière standard dans la programmation d'ordinateurs quantiques.

L'un des frameworks sort du lot, tout du moins côté usage chez les start-ups, est **TensorFlow** dont le développement a été initialisé par Google. Il fonctionne en embarqué aussi bien que sur serveurs et dans le cloud[135]. C'est le framework avec le spectre fonctionnel qui semble le plus large, et qui se déploie facilement sur des architectures parallèles, et notamment celles qui sont à base de GPU comme ceux de Nvidia.

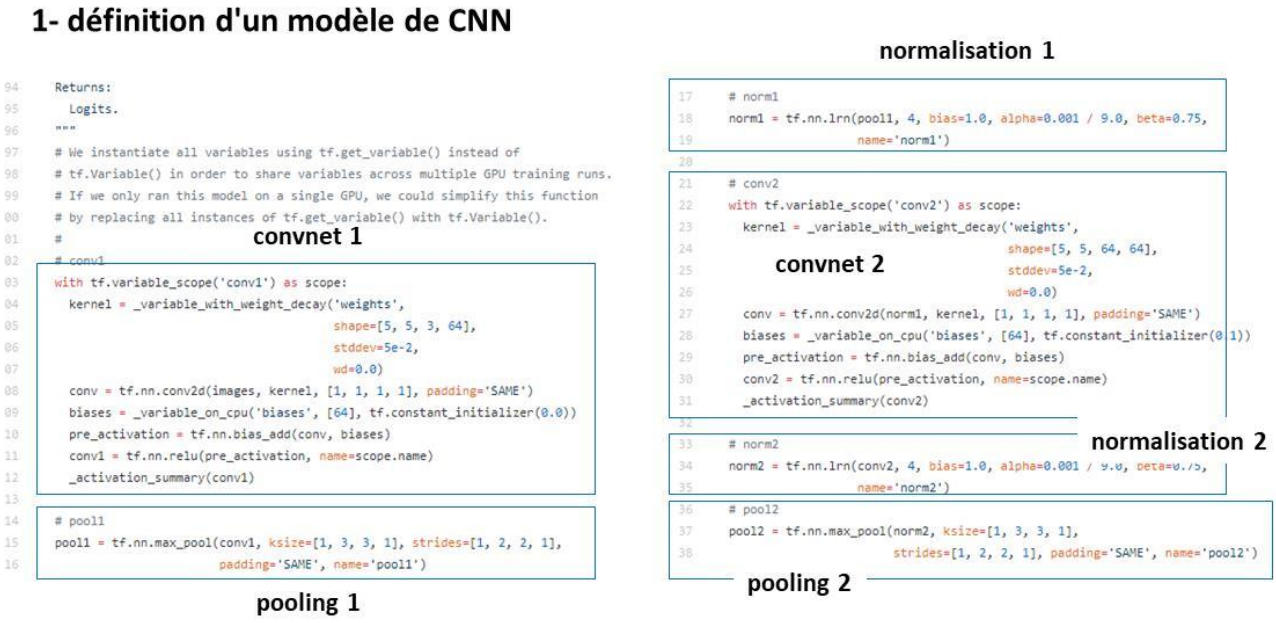
Cela explique qu'il ressorte du lot dans la petite comparaison *ci-contre* (source). Le nom TensorFlow vient de Tensor qui décrit les matrices multi-dimentionnelles gérées par le système. Google a annoncé au Google I/O de juin 2017 la sortie de TensorFlow Lite, une version allégée dédiée aux smartphones tournant sous Android. On peut imaginer qu'elle sera utilisable dans d'autres objets voire sur d'autres systèmes d'exploitation.



	Languages	Tutorials and training materials	CNN modeling capability	RNN modeling capability	Architecture: easy-to-use and modular front end	Speed	Multiple GPU support	Keras compatible
Theano	Python, C++	++	++	++	+	++	+	+
Tensor-Flow	Python	+++	+++	++	+++	++	++	+
Torch	Lua, Python (new)	+	+++	++	++	+++	++	
Caffe	C++	+	++		+	+	+	
MXNet	R, Python, Julia, Scala	++	++	+	++	++	+++	
Neon	Python	+	++	+	+	++	+	
CNTK	C++	+	+	+++	+	++	+	

TensorFlow est apprécié des startups car c'est le plus généraliste des frameworks.

Une autre solution populaire est **PyTorch**, un surensemble de Torch exploitable en Python. Alors que les données sont définies de manière statique dans Tensorflow, elles le sont de manière dynamique dans PyTorch, apportant une plus grande souplesse dans le développement.




```
  Returns:
    Logits.
  """
  # We instantiate all variables using tf.get_variable() instead of
  # tf.Variable() in order to share variables across multiple GPU training runs.
  # If we only ran this model on a single GPU, we could simplify this function
  # by replacing all instances of tf.get_variable() with tf.Variable().
  #
  # conv1
  with tf.variable_scope('conv1') as scope:
    kernel = _variable_with_weight_decay('weights',
                                         shape=[5, 5, 3, 64],
                                         stddev=5e-2,
                                         wd=0.0)
    conv = tf.nn.conv2d(images, kernel, [1, 1, 1, 1], padding='SAME')
    biases = _variable_on_cpu('biases', [64], tf.constant_initializer(0.0))
    pre_activation = tf.nn.bias_add(conv, biases)
    conv1 = tf.nn.relu(pre_activation, name=scope.name)
    _activation_summary(conv1)

  # pool1
  pool1 = tf.nn.max_pool(conv1, ksize=[1, 3, 3, 1], strides=[1, 2, 2, 1],
                         padding='SAME', name='pool1')
```

```
  # norm1
  norm1 = tf.nn.lrn(pool1, 4, bias=1.0, alpha=0.001 / 9.0, beta=0.75,
                    name='norm1')

  # conv2
  with tf.variable_scope('conv2') as scope:
    kernel = _variable_with_weight_decay('weights',
                                         shape=[5, 5, 64, 64],
                                         stddev=5e-2,
                                         wd=0.0)
    conv = tf.nn.conv2d(norm1, kernel, [1, 1, 1, 1], padding='SAME')
    biases = _variable_on_cpu('biases', [64], tf.constant_initializer(0.1))
    pre_activation = tf.nn.bias_add(conv, biases)
    conv2 = tf.nn.relu(pre_activation, name=scope.name)
    _activation_summary(conv2)

  # norm2
  norm2 = tf.nn.lrn(conv2, 4, bias=1.0, alpha=0.001 / 9.0, beta=0.75,
                    name='norm2')
  # pool2
  pool2 = tf.nn.max_pool(norm2, ksize=[1, 3, 3, 1],
                         strides=[1, 2, 2, 1], padding='SAME', name='pool2')
```

[135] Voici une source récente qui indique la popularité des frameworks d'IA : AI frameworks and hardware : who is using what?, mai 2018.

D'un point de vue pratique, à haut niveau, la programmation d'un réseau de neurones de deep learning revient à définir le modèle du réseau lui-même en décrivant de manière quasiment littérale une à une toutes ses couches (ci-dessous, un exemple en Tensorflow et Python).

Il faut ensuite programmer son entraînement, ce qui peut requérir de l'optimisation programmatique, puis son exécution en mode run-time (*ci-dessous*, toujours avec du code Python utilisant Tensorflow).

L'optimisation d'un réseau de neurones peut dépendre des capacités de l'architecture matérielle exploitée. Ainsi, la taille des filtres dans les réseaux convolutionnels pourra être liée à celle des multiplicateurs de matrices des GPU ou processeurs neuromorphiques utilisés dans les serveurs d'entraînement.

2- entrainement

propagation avant et gradients

```
def train(total_loss, global_step):
  """Train CIFAR-10 model.

  Create an optimizer and apply to all trainable variables. Add moving
  average for all trainable variables.

  Args:
    total_loss: Total loss from loss().
    global_step: Integer Variable counting the number of training steps
      processed.
  Returns:
    train_op: op for training.
  """
  # Variables that affect learning rate.
  num_batches_per_epoch = NUM_EXAMPLES_PER_EPOCH_FOR_TRAIN / FLAGS.batch_size
  decay_steps = int(num_batches_per_epoch * NUM_EPOCHS_PER_DECAY)

  # Decay the learning rate exponentially based on the number of steps.
  lr = tf.train.exponential_decay(INITIAL_LEARNING_RATE,
                                  global_step,
                                  decay_steps,
                                  LEARNING_RATE_DECAY_FACTOR,
                                  staircase=True)
  tf.summary.scalar('learning_rate', lr)

  # Generate moving averages of all losses and associated summaries.
  loss_averages_op = _add_loss_summaries(total_loss)
```

```
  # Compute gradients.
  with tf.control_dependencies([loss_averages_op]):
    opt = tf.train.GradientDescentOptimizer(lr)
    grads = opt.compute_gradients(total_loss)

  # Apply gradients.
  apply_gradient_op = opt.apply_gradients(grads, global_step=global_step)

  # Add histograms for trainable variables.
  for var in tf.trainable_variables():
    tf.summary.histogram(var.op.name, var)

  # Add histograms for gradients.
  for grad, var in grads:
    if grad is not None:
      tf.summary.histogram(var.op.name + '/gradients', grad)

  # Track the moving averages of all trainable variables.
  variable_averages = tf.train.ExponentialMovingAverage(
      MOVING_AVERAGE_DECAY, global_step)
  variables_averages_op = variable_averages.apply(tf.trainable_variables())

  with tf.control_dependencies([apply_gradient_op, variables_averages_op]):
    train_op = tf.no_op(name='train')

  return train_op
```

De son côté, **Theano** est un projet académique lancé par l'Université de Montréal. Il était initialement très bien supporté et apprécié pour sa rapidité de fonctionnement. Il est aussi assez couramment utilisé dans les startups jusqu'en 2016. Mais TensorFlow a pris le dessus depuis 2016 à une vitesse incroyable et l'équipe de Theano a annoncé qu'il ne serait plus supporté à partir de 2018. Un retournement de marché en moins de deux ans ! Cela montre le côté impitoyable des batailles de plateformes et la difficulté que peuvent avoir des laboratoires de recherche à l'origine de projets open source lorsqu'ils sont face à de grands acteurs du numérique avec leur force marketing. C'est une leçon à retenir pour les laboratoires français, comme l'INRIA et son framework **scikit-learn** qui jusqu'ici résiste plutôt bien.

Certains outils sont exploités de manière combinée. Ainsi, la bibliothèque de prototypage de deep learning **Keras**, créé par le Français François Chollet (qui travaille chez Google depuis 2015), peut-elle s'appuyer sur **TensorFlow**.

Ces différents outils sont aussi disponibles dans des offres en cloud, notamment chez Google, Amazon, Microsoft, IBM et même chez OVH.

solutions de deep learning dans le cloud

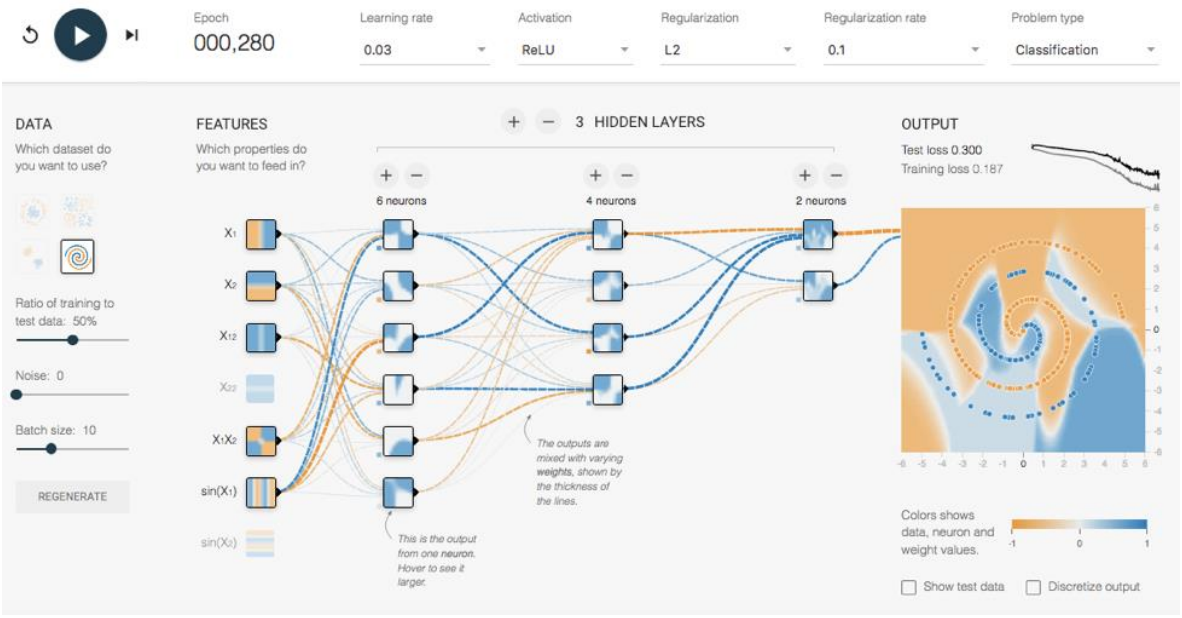
Google	**Google Cloud Machine Learning**
amazon	**Amazon Artificial Intelligence & Alexa**
Microsoft	**Microsoft Cognitive Services + Azure**
IBM	**IBM Watson Cloud Servers**
OVH.com	**hébergement de serveurs Nvidia**

Voici quelques-unes des solutions de deep learning les plus courantes pour les développeurs de solutions extraites de l'édition 2018 du Guide des Startups.

Outil	Usage
	TensorFlow est une bibliothèque open source de développement d'applications de machine learning déployable dans le cloud de manière répartie ainsi que dans l'embarqué. Elle est proposée sous forme de service en cloud par Google. Elle sert notamment à détecter des patterns, à faire de la classification automatique et des prévisions. Les Tensor Processing Units sont des processeurs dédiés au traitement avec TensorFlow qui ont été développés par Google pour son offre en cloud. Ils ont été notamment utilisés pour faire gagner DeepMind au jeu de Go en 2016. En 2017, les TPU en étaient déjà à leur seconde génération. Les applications TensorFlow sont faciles à déployer sur des architectures réparties sur plusieurs CPU et GPU.
	scikit-learn est un kit de développement d'applications de machine learning et de deep learning mettant en œuvre les méthodes de classification, de régression (prévisions) et de clustering. Il s'exploite en Python. La solution est en open source sous license BSD et est issue de l'INRIA et de Telecom Paritech. Sa communauté internationale comprend 1135 contributeurs depuis sa création avec environ 70 actifs par version. La documentation de scikit-learn fait 2373 pages (release 0.20.0, 27 septembre 2018) !
	Keras est une bibliothèque open source écrite en Python qui s'appuie sur Deeplearning4j, Tensorflow ou Theano. Elle permet de créer des solutions de deep learning avec un plus haut niveau d'abstraction que TensorFlow. Elle est issue du projet ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), et son principal auteur et contributeur est un français, François Chollet, qui travaille chez Google.
	CNTK est un framework open source de Deep Learning de Microsoft qui fait partie de leur Cognitive Toolkit qui permet notamment de créer des agents conversationnels (chatbots). Microsoft propose une gamme d'API complète pour presque toutes les applications de machine learning et de deep learning.
	IBM Watson est la solution d'intelligence artificielle d'IBM. C'est en fait un ensemble de briques logicielles assez complet permettant de bâtir différentes formes d'applications d'intelligence artificielle, avec ce qui concerne la reconnaissance des images, de la parole ou de textes ainsi que l'automatisation du raisonnement avec des moteurs de règles et des solveurs. La solution a des usages multiples : création de robots conversationnels, aide au diagnostic et à la prescription dans l'imagerie médicale, prévisions dans la finance, aide juridique automatisée, cybersécurité, etc. Watson est notamment fourni en cloud. Il est assez couramment utilisé par les startups, IBM étant très actif dans leur recrutement.
	MxNet est une bibliothèque supportée par Amazon supportant le développement en Python, C, C++ et Julia. Il est distribuable sur plusieurs GPU et CPU.
	Clarifai est une solution de deep learning en cloud qui sert notamment à la reconnaissance d'images, en particulier dans la santé et pour la création de moteurs de recherche d'images.
	PaddlePaddle est la bibliothèque de deep learning de Baidu adaptée aux traitements distribués. Elle est fournie avec un outil de visualisation de l'état de l'entraînement de son réseaux de neurones.
	PyBrain est une bibliothèque de réseau de neurones bâtie en Python.
	Caret (Classification And REgression Training) est une bibliothèque permettant de développer des applications de prévisions.
	Vowpal Wabbit est une bibliothèque open source provenant de Yahoo! Research et gérée par Microsoft Research. Elle permet de créer des solutions de machine learning en ligne. Comme la grande majorité des bibliothèques de machine learning, elle sert à faire de la classification automatique, de la prévision, de la segmentation. Elle exploite des CPU multi cœurs et est compilée à partir de C++.

Caffe2	Caffe2 est un framework open source générique de deep learning. La première version provenait du Berkeley Vision and Learning Center et avait été développée avec l'aide financière de Nvidia et Amazon. La seconde a bénéficié de la contribution de Facebook. Le framework sait notamment exploiter les serveurs GPU Nvidia et en mode distribué. Les réseaux entraînés sont aussi facilement déployables sur architectures mobiles. On utilise Caffe en Python ou C++. Le framework facilite la distribution de traitement sur plusieurs processeurs.
torch PYTORCH	Torch est un framework de deep learning utilisé notamment dans la vision artificielle. Il est utilisé chez Facebook, Google et Twitter et provient de l'Université de New York. On l'exploite notamment avec le langage Lua qui est une sorte de Python simplifié. C'est le framework préféré de Yann LeCun ! A noter la déclinaison PyTorch qui est exploitable avec Python, un langage plus populaire. PyTorch est apprécié pour le prototypage de solutions de deep learning. PyTorch permet de générer dynamiquement les graphes des réseaux de neurones programmés au moment de l'exécution du code ce qui facilite leur mise au point.
H2O.ai	H2O.ai est un framework open source de machine et deep learning couramment utilisé par les data scientists. La startup qui en est à l'origine a levé $33,6M. Elle est associée à un backend de distribution de traitements (Map/Reduce). Elle est exploitable à partir de nombreux langages comme R, Python, Java et Scala et via des API REST. Au passage, au printemps 2017, H2O, Continuum Analytics et MapD Technologies lancaient l'initiative GPU Open Analytics Initiative (GOAI) pour créer un framework ouvert commun destiné à l'exploitation en mémoire d'analytics sur GPU. Le tout avec la bénédiction de Nvidia.
spaCy	spaCy est une bibliothèque open source de traitement du langage pour Python. Elle permet d'analyser rapidement le contenu de textes en anglais, français, allemand et espagnol. Elle s'interface avec TensorFlow. Keras et scikit-learn.
mahout	Originaire de la fondation Apache, Mahout est un framework qui permet de développer des applications d'IA scalable, en particulier dans des applications de classification automatique et de filtrage collaboratif. Il est utilisé chez Amazon.
Spark MLlib	Egalement originaire de la fondation Apache, MLlib est une bibliothèque de machine learning pour Apache Spark, un système de répartition de traitements concurrent de Hadoop. On l'exploite en Java, Scala, Python et en R. Elle servira à traiter de gros volumes de données pour les exploiter avec des algorithmes de classification, régression ou segmentation automatique.
opennn neural networks	OpenNN est une classe open source en C++ de la startup espagnole Artelnics qui sert à développer des applications de réseaux neuronaux.
PredictionIO	PredictionIO est une solution open source pour développer des applications de machine learning dédiées à la prévision. C'est une sorte de "MySQL pour le machine learning". C'est encore une solution de la fondation Apache.
ALGORITHMIA	Algorithmia est une place de marché d'algorithmes et de briques logicielles d'IA qui sont positionnées comme des « micro services », et disponibles en cloud, faciles à tester, intégrer et mettre en production. Les services proposés sont assez classiques comme la détection de visage dans des photos ou l'analyse de sentiments dans les flux de réseaux sociaux.
Stanford University S NLP Natural Language Processing	Stanford CoreNLP est un framework généraliste de traitement du langage qui permet d'identifier la structure grammaticale et syntactique de phrases, de détecter des mots clés, des nombres, analyse des sentiments, etc. Il fonctionne dans de nombreux langages. Les fonctions sont exploitables via les principaux langages du marché ainsi qu'au travers de services web. Le framework est open source.
OPEN NLP	Open NLP est un framework de traitement du langage de la fondation Apache créé en 2004 et qui a évidemment bien évolué depuis. Il comprend les fonctions classiques de décomposition de phrases en entités, d'analyse de structure grammaticale et syntaxique et d'extraction de données.

Vous pouvez aussi simuler des réseaux de neurones simples avec cet outil exploitant **TensorFlow** (exemple *ci-dessous*). TensorFlow peut en effet servir aussi bien à gérer un réseau de neurones multicouches de deep learning tout comme des solutions plus simples de machine learning, à base ou pas de réseaux de neurones simples.



Vous pouvez télécharger ces deux ouvrages : Deep Learning in Python de François Cholet 2018 (386 pages) et Deep learning with Keras 2017 (490 pages). Et aussi reproduire chez vous ces 30 tests d'applications de deep learning qui sont en open source[136].

Citons une dernière catégorie d'outils, les places de marché de modèles d'IA pour les développeurs. En voici quelques-unes :

- **ONNX** (2016, USA) est un format de modèles de deep learning et machine learning interopérable supporté par Caffe2, Chainer, MxNet, PyTorch, PaddlePaddle, Microsoft mais pas TensorFlow. Peut-être parce que c'est un projet communautaire poussé par Amazon, Facebook et Microsoft. La solution permet l'import et l'export de modèles de et vers ces différents outils.
- **SingularityNET** (2017, Suisse) est une plateforme de monétisation de services d'IA. C'est une sorte de version commerciale d'ONNX. Elle a éé cofondée par le fameux Ben Goertzel, un spécialiste de l'AGI. Mais là, ce n'est pas de l'AGI. Mais il en parle dans Toward Grand Unified AGI[137].
- **Algorithmia** (2013, USA, $12,9M) est autre place de marché d'algorithmes d'IA. Elle propose à ce jour 5000 algorithmes, probablement pas tous du domaine de l'IA.

En marge des outils de création de solutions à base de deep learning, on peut signaler que le machine learning et le deep learning sont aussi mis au service des développeurs dans le cadre d'outils de tests de code, souvent mis en œuvre via des solutions en cloud.

En voici quelques-uns :

- **DeepCode** (2016, Suisse, $700K) qui analyse le code pour l'améliorer à partir d'un corpus de 250 000 règles de bonne programmation. Il va notamment consulter le code que vous avez publié sur Github. Ils proposent aussi JSNice, qui *déobsfuscate* du code JavaScript minifié. C'est une spin-off de l'Université ETH Zurich.

[136] Voir 30 Amazing Machine Learning Projects for the Past Year (v.2018), janvier 2018.

[137] Voir le livre blanc SingularityNET: A decentralized, open market and inter-network for AIs, décembre 2017 (53 pages).

- **Functionize** (2015, USA, $3,2M) propose un outil de tests d'applications utilisateurs permettant d'exprimer un parcours visiteur en langage naturel, qui est ensuite analysé par l'outil et exécuté via un plugin tournant dans le navigateur Chrome. L'outil génère aussi des indicateurs de performance du code.
- **Applitools** (2013, Israël, $41,8M) réalise comme Functionize des tests fonctionnels d'applications web. Il utilise de la reconnaissance d'image pour analyser les écrans et identifier des variations ou régressions en ne tenant compte que des éléments perceptibles par l'œil humain.
- **Testim** (2014, USA, $7,9M) propose un outil de gestion de tests de logiciels. Il s'appuie sur du machine learning pour accélérer la création, l'exécution et la maintenance de tests automatisés.
- **Sealights** (2015, Israël, $11M) utilise du machine learning pour analyser le code à tester et les tests qui y sont appliqués, qu'il s'agisse de tests fonctionnels ou de performance.
- **Test.AI** (2015, USA, $17,6M) utilise un outil de detection automatique des composantes visuelles d'une application pour générer des tests. Cela semble surtout cibler les applications mobiles.
- **Mabl** (2017, USA, $10M) utilise le machine learning pour gérer les tests et identifier les régressions dans les évolutions du code. L'outil adapte automatiquement les tests en fonction de la détection d'évolutions de l'interface utilisateur de l'application. La société a été créée par des anciens de Google.

Va-t-on pour autant voir apparaître des outils de développement à base d'IA qui vont remplacer les développeurs ? C'est une thèse partagée par quelques prospectivistes de l'IA. Je n'y souscris pas. Créer un logiciel nécessite toujours de formaliser le besoin. Que ce soit de manière graphique ou par du code, on doit toujours en passer par là. Ce qui pourra arriver est la prolifération de solutions permettant d'exploiter des briques d'IA à haut niveau sans avoir à connaître les dessous du machine learning et du deep learning. Des évolutions des outils de développement et de tests à base d'IA feront cependant leur apparition qui assisteront les développeurs. Mais pas de là à les remplacer.

IA explicable

L'une des craintes associées à l'IA connexionniste en général et au deep learning en particulier est de générer des solutions logicielles dont le fonctionnement n'est pas facilement explicable, interprétable et auditable. Les exemples les plus faciles à visualiser sont ces réseaux de neurones convolutionnels qui sont capables de caractériser des images.

Ils s'appuient sur des filtres dans plusieurs couches de convolution qui ne sont pas créés par l'Homme mais automatiquement générés par le mécanisme de la rétropropagation pendant l'entraînement.

Ces filtres correspondent à une analyse probabiliste automatique permettant au réseau de neurones de distinguer des formes de niveaux d'abstraction empilés. Mais ces niveaux d'abstraction ne correspondent pas forcément à ceux de la vision humaine. Donc, on a du mal à comprendre le cheminement de l'information lorsque l'on traverse les différentes couches du réseau de neurones.

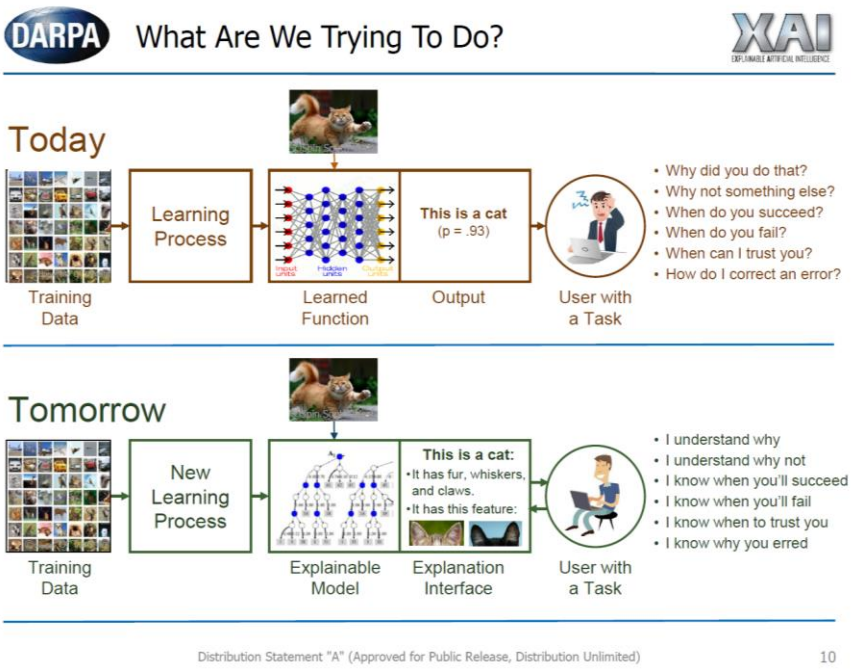
Il en va de même pour l'analyse du langage pour détecter des sentiments, réaliser des traductions ou extraire des données. Les modèles probabilistes du deep learning des réseaux de neurones de traitement du langage, mêlant souvent des convnets, des RNN (réseaux récurrents) et des LSTM (réseaux récurrents à mémoire), étant générés automatiquement, ils sont difficiles à auditer.

Difficile, mais pas impossible. La notion d'IA explicable est devenue un champ scientifique à part entière de l'IA. On l'appelle XAI pour Explainable AI[138]. C'est devenu une priorité presque partout dans le monde, dans la recherche aux USA, en Chine et en France.

L'agence DARPA du Pentagone a lancé sa propre initiative XAI dédiée en 2017.

Dans le schéma *ci-contre*, on comprend bien l'objectif : être capable de décomposer les étapes qui mènent un réseau convolutif à reconnaître un objet (ici, un chat) dans une image.

L'intention est également de mieux comprendre les biais éventuels des réseaux de neurones pour identifier les sources d'erreurs potentielles. Il faut noter une distinction entre l'explicabilité et l'interprétabilité.



L'**explicabilité** concerne en gros le processus et l'algorithme en général et l'**interprétabilité** décrit le lien entre les solutions et les données d'entraînement, à savoir quelles sont les données qui ont influencé le résultat fourni par l'IA.

Les questions que l'utilisateur peut se poser concernant le résultat d'une IA sont : Pourquoi ce résultat a-t-il été obtenu ? Pourquoi une autre méthode n'a-t-elle pas été mise en œuvre ? Dans quelles conditions la méthode retenue fonctionne-t-elle ? Quand échoue-t-elle ? Dans quelles conditions puis-je lui faire confiance ? Comment savoir si une erreur a été commise et comment la corriger ? On devrait aussi ajouter : quelles sont les données qui ont servi à entraîner le modèle de l'IA [139]? Et ces données sont-elles adaptées au problème ?

Les méthodes rendant les IA explicables sont aussi nombreuses que les méthodes d'IA elles-mêmes[140].

Dans nombre de cas de figure, ce sont souvent des questions plus académiques que pratiques. Elles sont surtout pertinentes pour gérer les cas les plus épineux comme un audit après un accident mettant en œuvre un véhicule à conduite autonome ou l'aide à la prise de décision dans les questions médicales, surtout de vie ou de mort. Les juristes sont évidemment très sensibles à ces questions pour à la fois faire évoluer le droit et leurs pratiques[141].

---

[138] Voir les actes du workshop sur l'XAI de la conférence IJCAI de 2017, l'une des plus importantes de l'IA au monde, dans IJCAI-17 Workshop on Explainable AI (XAI), 2017 (68 pages). Et cet article de vulgarisation sur la question : The Dark Secret at the Heart of AI de Will Knight, MIT Technology Review, avril 2017.

[139] Dans Increasing Trust in AI Services through Supplier's Declarations of Conformity, août 2018 (29 pages), une brochette de chercheurs d'IBM Research propose la création d'un cahier des charges d'IA explicables. Ils ajoutent aux questions de base les questions suivantes : quel est l'usage du service ? Quels algorithmes et techniques sont employés ? Décrire les méthodes de tests et leurs résultats. Existe-t-il des biais connus ayant une incidence éthique ou en matière de sécurité ? Quels efforts ont été menés pour les éviter ? Est-ce que le service a été vérifié contre les attaques de GANs (Generative Adversarial Networks) ? Quand le modèle a été mis à jour la dernière fois ?

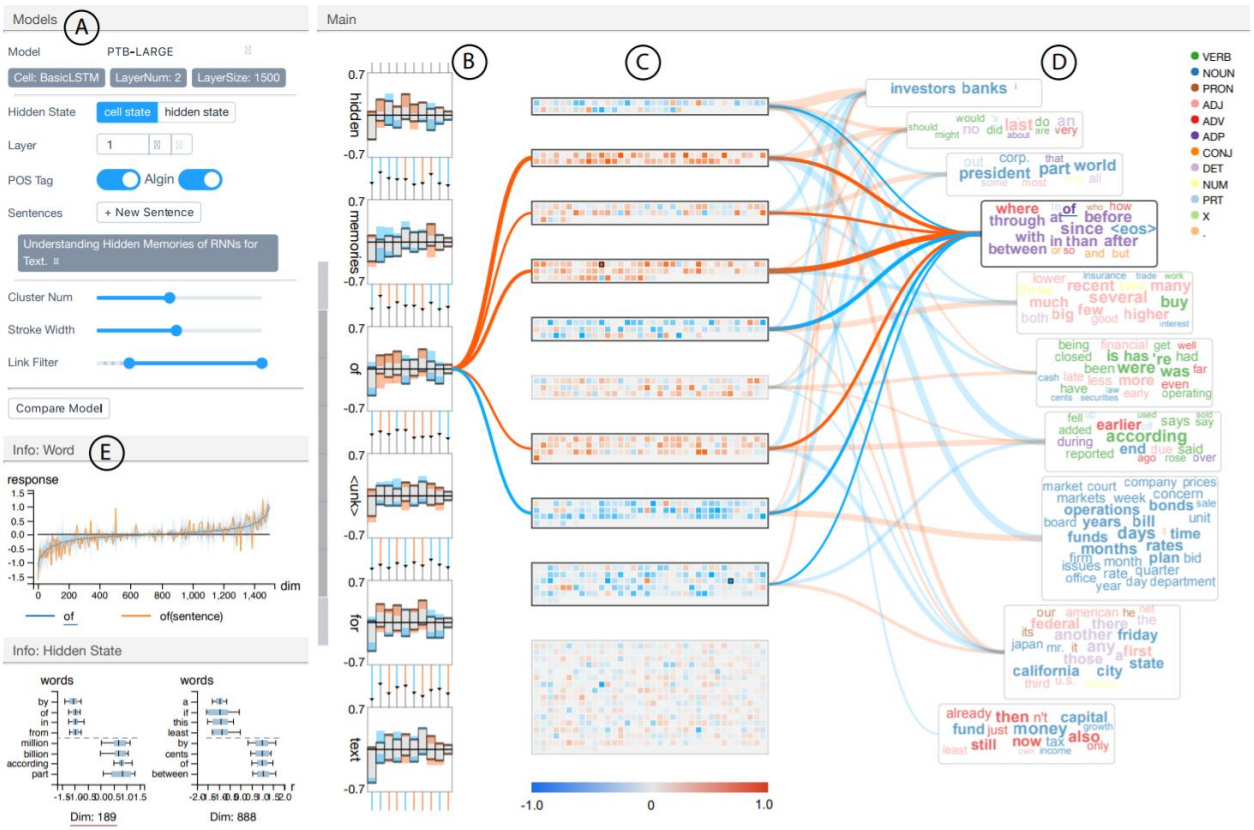
[140] Pour ce qui est de la reconnaissance d'images, quelques méthodes de XAI sont très bien documentées dans Towards explainable Deep Learning de Wojciech Samek du Fraunhofer HHI, novembre 2017 (88 slides).

[141] Voir Accountability of AI Under the Law: The Role of Explanation, de Finale Doshi-Velez et Mason Kortz, 2017.

Nombre de logiciels prennent des décisions pour nous sans que l'on puisse les auditer, même lorsqu'ils n'utilisent pas forcément de l'IA. Les exemples classiques concernent le ciblage publicitaire, les flux d'information dans les réseaux sociaux ou la décision de mener un contrôle fiscal sur vous ou votre entreprise ! Néanmoins, une IA non explicable utilisée seule n'est pas compatible avec le RGPD[142].

Mais il y a des cas bien pratiques où une explication serait la bienvenue. Ainsi, j'ai voulu poster un commentaire sur un produit défectueux acheté sur Amazon en août 2018.

Amazon m'a répondu que mon commentaire d'une ligne n'était pas conforme aux règles en vigueur chez eux. Mais en oubliant de préciser ce qui n'allait pas dans mon texte et la règle violée. Ce n'était pas du tout évident en consultant lesdites règles qui sont étalées sur plusieurs pages. Même s'il est probable que le « *It's a joke* » n'était pas acceptable pour eux[143], ce qui est pour le moins étonnant.

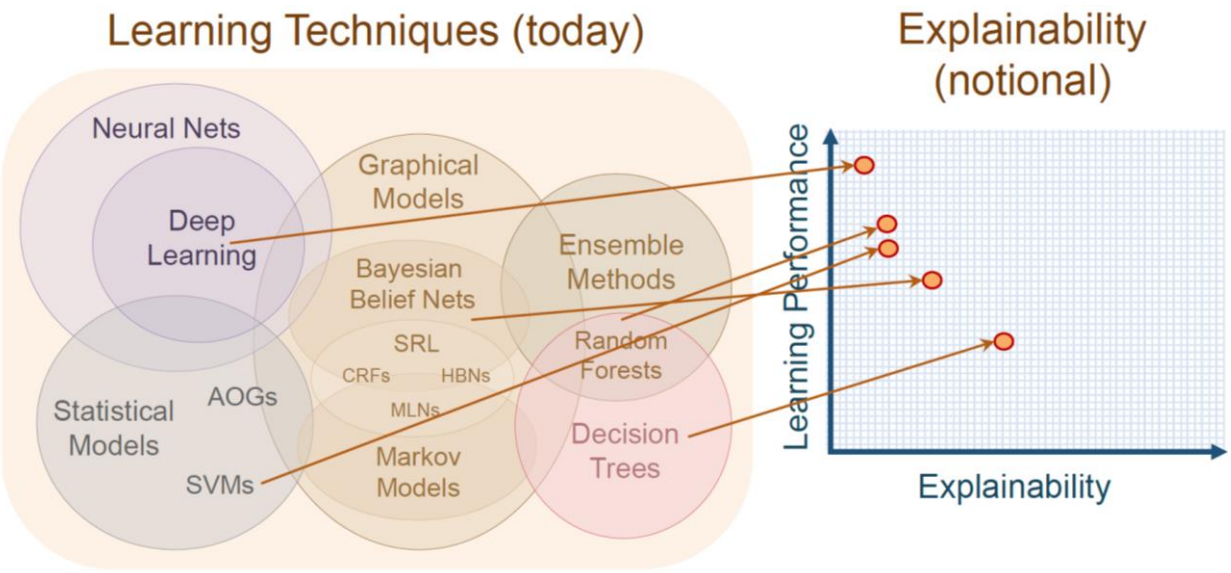


[142] Voir GDPR / RGPD - Protection des données incompatible avec Deep Learning de Dimitri Carbonnelle Livosphere 2017. Selon son article 63 « *toute personne concernée devrait avoir le droit de connaître et de se faire communiquer [...] les finalités du traitement des données à caractère personnel, [...] la logique qui sous-tend leur éventuel traitement automatisé et les conséquences que ce traitement pourrait avoir, au moins en cas de profilage.* ». Une banque ne peut ainsi pas utiliser une solution à base de deep learning pour évaluer la solvabilité d'un client potentiel et lui refuser un prêt sans autre explication. Elle peut exploiter la solution mais doit expliquer au client pourquoi elle refuse ce prêt. Le deep learning est un outil d'aide à la décision et pas un outil de prise de décision autonome.

[143] C'est l'objet de travaux comme Understanding Hidden Memories of Recurrent Neural Networks, 2017 (16 pages) qui vise à améliorer l'explicabilité des réseaux de neurones récurrents dans le traitement du langage.

Cela amène la notion d'interface d'IA explicable, présentée par la DARPA. Il ne suffit pas de pouvoir expliquer le résultat d'un traitement, il faut pouvoir en visualiser l'explication de manière compréhensible pour les utilisateurs. Cela peut relever d'approches visuelles ou d'explications en langage naturel.

L'un des défis de l'Explainable AI est de bien équilibrer la recherche d'explicabilité et la performance des techniques utilisées. Comme présenté dans le schéma *ci-dessous*, le paradoxe du deep learning est d'être très performant mais faiblement explicable tandis que d'autres méthodes sont au contraire explicables mais moins performantes. L'idée est de conserver la bonne performance du deep learning tout en améliorant son explicabilité.



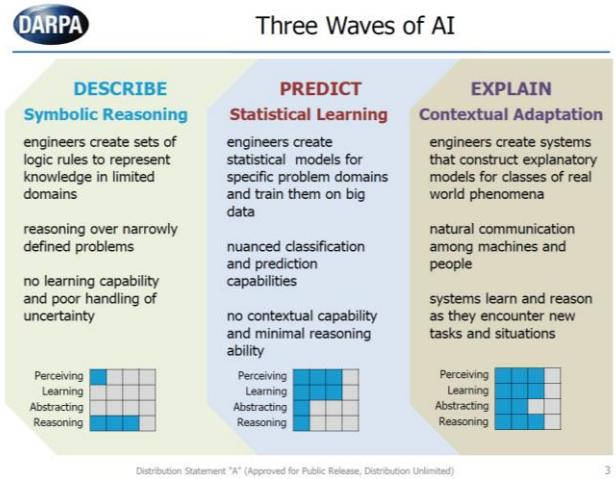
L'IA explicable doit aussi répondre à des exigences d'éthique (biais, …), de confiance (équité, rigueur, …) et de sécurité (respect de la vie privée, non détournement des données personnelles pour des usages non documentés, …)[144].

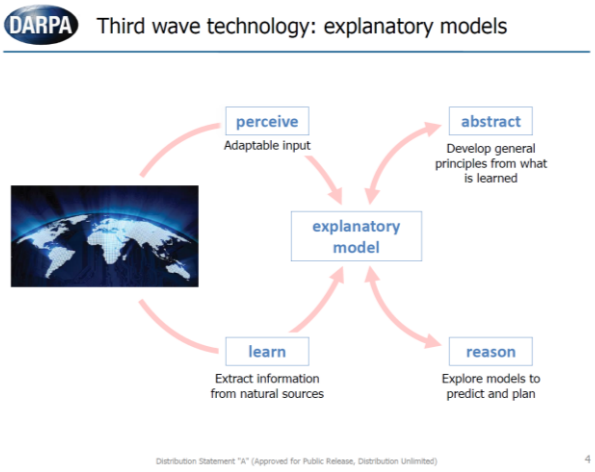
David Gunning de la DARPA explique très bien toute la démarche dans Explainable Artificial Intelligence (XAI) Program Update, novembre 2017 (36 slides). Cette initiative a sélectionné et financé divers projets dans des universités telles que Berkeley, UCLA, Oregon State et Carnegie Mellon ainsi que dans des laboratoires de recherche privés comme le PARC de Xerox[145], SRI International et Raytheon BBN.

Comme l'IA washing est courante, on risque de voir émerger de l'XAI washing ou tout du moins un marketing de l'offre mettant bien l'accent sur l'explicabilité de l'IA. Cela commence avec **Maathics** (France), une startup qui propose un système d'audit automatisé d'algorithmes d'IA permettant « *un usage équitable de l'IA* ». Reste à savoir comment cela fonctionne et si le système est lui-même explicable !

[144] Voir Vers une intelligence artificielle ubiquitaire explicable du chercheur en cybersécurité Thierry Berthier, juillet 2018 ainsi que Explanation in Artificial Intelligence: Insights from the Social Sciences, de Tim Miller, 2018 (66 pages).

[145] Pour le projet COGLE (COmmon Ground Learning and Explanation), voir PARC to Develop Explainable Artificial Intelligence (XAI) Science for DARPA, juillet 2017. Le projet vise à rendre explicable les décisions de systèmes autonomes tels que des voitures autonomes ou des drones.





Agents et réseaux d'agents

Dans ce concept apparu dans les années 1990, les agents intelligents permettent de résoudre des problèmes dans des architectures distribuées. Conceptuellement, un agent est un logiciel ou un matériel qui capte de l'information, décide d'agir rationnellement en fonction des données récupérées et déclenche une action pour optimiser ses chances de succès.

Si c'est du matériel, il comprendra des capteurs et des actuateurs. Mais il peut n'être que du logiciel et obtenir des données brutes en entrées et générer des données en sortie.

Un agent réagit donc en fonction de l'environnement et de préférence en temps réel. Les agents intelligents sont intégrés dans des systèmes distribués dénommés systèmes multi-agents avec des agents autonomes, mais reliés et collaborant entre eux.

Les agents sont autonomes, ils appliquent des règles et vont jusqu'à apprendre à les modifier en fonction de l'environnement, ils peuvent être proactifs et pas seulement réactifs à l'environnement, ils communiquent et coopèrent avec d'autres agents et systèmes.

Les solutions d'intelligence artificielle sont ainsi souvent des réseaux multi-agents ! C'est notamment le cas des solutions de RPA (Robotic Process Automation) utilisées dans le backoffice des banques et des assurances.

Les réseaux d'agents fonctionnent de manière coordonnée et collective. La coordination de réseaux d'agents est un domaine scientifique à part entière.

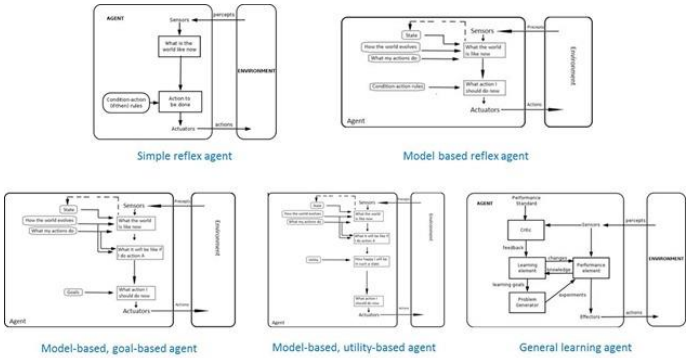
On compte notamment les **Distributed Problem Solving** (DPS) qui découpent un problème en sous-problèmes qui sont résolus de manière coopérative entre plusieurs agents reliés les uns aux autres. Ces systèmes sont conçus pour résoudre des problèmes bien spécifiques.

Les types d'agents

Les agents sont classifiés par Russell & Norvig dans **Artificial Intelligence – A Modern Approach** (2003-2009) en **types distincts** selon leur niveau d'autonomie et leur mode de prise de décision :

- Les **simple reflex agents** qui comprennent des capteurs, des règles indiquant quelle action mener et des actuateurs pour les déclencher. Ils travaillent en temps réel.
- Les **model based reflex agents** qui ajoutent un moteur d'état capable de mémoriser dans quel état se trouve l'objet et qui évaluent l'impact des actions pour changer d'état.
- Les **goal-based agents** qui prennent leur décision en fonction d'un objectif et déterminent une action pour l'atteindre.

- Les **utility-based agents** qui prennent leur décision en fonction d'un but à atteindre qui est plus général.
- Les **learning agents** qui contiennent une fonction d'auto-apprentissage.



intégration et réseaux d'agents

agent = système qui réagit à son environnement
- homme, animal, robot, chatbot, système d'IA, logiciel
- capteurs + outils d'action sur l'environnement
- contexte => action => réaction => évaluation

solutions d'IA = agents ou réseaux d'agents
- agent conversationnel
- robot aspirateur
- groupe de robots ou de drones coordonnés
- personnages virtuels dans un jeu

Vu de haut, les réseaux d'agents ressemblent aux réseaux de neurones mais leur mode de fonctionnement est différent. Un agent peut très bien être lui-même individuellement construit avec un réseau de neurones pour réaliser une tâche spécifique comme la reconnaissance de la parole ou d'images. Et la liaison entre les agents se fait selon des règles, comme dans un moteur de règles ou de workflow !

Les solutions d'IA sont des Systèmes Multi-Agents

Un autre agent va utiliser le texte généré par la reconnaissance puis appliquer un processus de reconnaissance sémantique, puis un autre va traiter la question, fouiller dans une base de données ou de connaissance, récupérer des résultats, un autre va formuler une réponse et la renvoyer à l'utilisateur. Idem pour un système de traduction automatique qui va d'abord analyser la parole avec un premier agent, puis réaliser la traduction avec un second, puis utiliser un troisième agent de "text to speech" pour transformer le résultat de manière audible.

Un robot conversationnel est aussi un réseau d'agents, surtout si on interagit avec la voix avec lui. Les agents sont notamment utilisés dans les systèmes de call centers. Une start-up française s'était lancée - parmi d'autres - sur ce créneau : **Virtuoz**. Elle a été acquise en 2013 par l'américain **Nuance**. Il existe même un concours du **meilleur agent de service client en ligne**, lancé en 2016 en France avec une trentaine de candidats !

Un robot autonome est aussi un condensé de nombreux agents qui gèrent différents niveaux d'abstraction avec de nombreux capteurs, de la mécanique, des systèmes permettant au robot de savoir où il est, avec quoi il interagit, et qui a des missions à accomplir (aider une personne, conduire un véhicule, etc).

Un robot est particulièrement complexe à mettre au point car il cumule des défis au niveau des capteurs, de l'intégration de ses sens, de la mécanique pour se mouvoir, de la batterie pour son autonomie, et dans l'intelligence artificielle pour piloter l'ensemble et éventuellement interagir à la fois mécaniquement, visuellement et oralement avec son environnement, notamment s'il s'agit de personnes.

C'est dans le domaine de l'**intelligence artificielle intégrative** que des progrès significatifs peuvent être réalisés. Elle consiste à associer différentes méthodes et techniques pour résoudre des problèmes complexes voire même résoudre des problèmes génériques. On la retrouve mise en œuvre dans les agents conversationnels tels que ceux que permet de créer IBM Watson ou ses concurrents.

Dans le jargon de l'innovation, on appelle cela de l'innovation par l'intégration. C'est d'ailleurs la forme la plus courante d'innovation et l'IA ne devrait pas y échapper. Cette innovation par l'intégration est d'autant plus pertinente que les solutions d'IA relèvent encore souvent de l'artisanat et nécessitent beaucoup d'expérimentation et d'ajustements.

Cette intégration est un savoir nouveau à forte valeur ajoutée, au-delà de l'intégration traditionnelle de logiciels via des APIs classiques. Cette intelligence artificielle intégrative est à l'œuvre dans un grand nombre de startups du secteur et en particulier dans celles de la robotique.

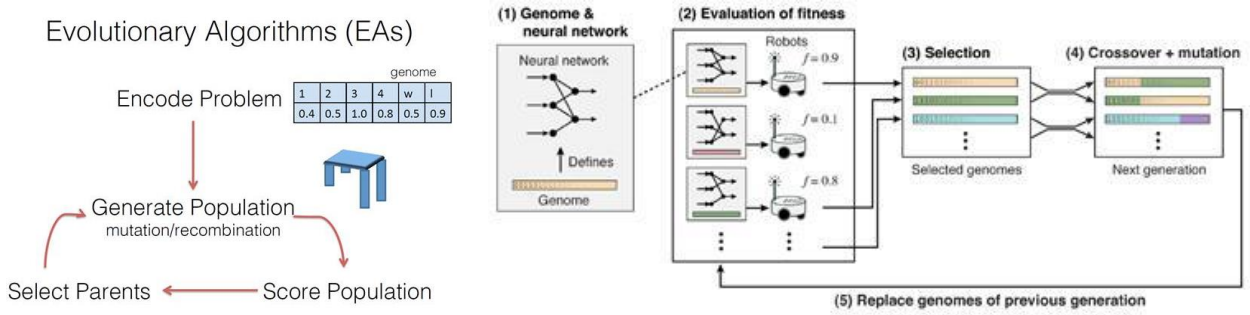
Le mélange des genres n'est pas évident à décrypter pour le profane : machine learning, deep learning, support vector machines, modèles de Markov, réseaux bayésiens, réseaux neuronaux, méthodes d'apprentissage supervisées ou non supervisées, etc. D'où une discipline qui est difficile à benchmarker d'un point de vue strictement technique et d'égal à égal. Ce d'autant plus que le marché étant très fragmenté, il y a peu de points de comparaison possibles entre solutions. Soit il s'agit de produits finis du grand public comme la reconnaissance d'images ou vocale, et d'agents conversationnels très à la mode en ce moment, soit il s'agit de solutions d'entreprises exploitant des jeux de données non publics.

Quid des outils de développement associés à la création de réseaux d'agents ? **Il y en a plein**, et notamment en open source, tels que **Soar** de l'Université du Michigan qui servent à mener des tâches et raisonnements généraux ou **REPAST**, de l'Université de Chicago, qui permet de créer des simulations d'interactions de millions d'agents[146]. On peut aussi intégrer **Unity Machine Learning Agents** qui servira notamment aux développeurs de jeux vidéo, très friands de réseaux d'agents ainsi que **JADE** (Java Agent DEvelopment framework) qui vient de Telecom Italia et est open source[147].

En France, **Yves Demaseau** du Laboratoire d'Informatique de Grenoble (CNRS) est un spécialiste des Systèmes **M**ulti-**A**gents. Il a monté un programme de recherche de 10 ans sur le sujet pour créer un système de SMA d'un million d'agents. Il est par ailleurs président de l'**Association Française pour l'Intelligence Artificielle**.

La programmation génétique

La **vie artificielle** et la **programmation génétique** sont d'autres pans de recherche important connexe aux recherches sur l'IA. Il s'agit de créer des modèles permettant de simuler la vie avec un niveau d'abstraction plus ou moins élevé. On peut ainsi simuler des comportements complexes intégrant des systèmes qui s'auto-organisent, s'auto-réparent, s'auto-répliquent et évoluent d'eux-mêmes en fonction de contraintes environnementales. Et les éléments les moins efficaces de ces systèmes sont éliminés, comme dans le processus de sélection naturelle décrit par Darwin.



Ces systèmes exploitent des algorithmes évolutifs qui sont à la croisée des chemins du deep learning et des réseaux d'agents. Ils consistent à tester différentes combinaisons de réseaux de neurones voire de réseaux d'agents les intégrant pour comparer leur efficacité et conserver les variantes les plus efficaces.

---

[146] La conférence de référence sur le sujet des agents semble être l'AAMAS, la dernière édition ayant eu lieu à Stockholm en avril 2018. Voir la liste des papiers publiés à cette occasion.

[147] Voir aussi le cours Modélisation et Simulation Multi-Agents - Cours 1 - Prolégomènes de Jean-Daniel Kant (103 slides) qui décrit bien la notion de SMA ainsi que Outils et langages de programmation d'applications multi-agents 2017 (46 slides) qui fait un inventaire actualisé des outils de développement de systèmes multi-agents.

C'est une reproduction informatique du principe de la sélection darwinienne. Reste à s'assurer qu'ils sont efficaces, ce qui est loin d'être évident vue la combinatoire de scénarios qu'ils peuvent être amenés à simuler !

Artificial General Intelligence

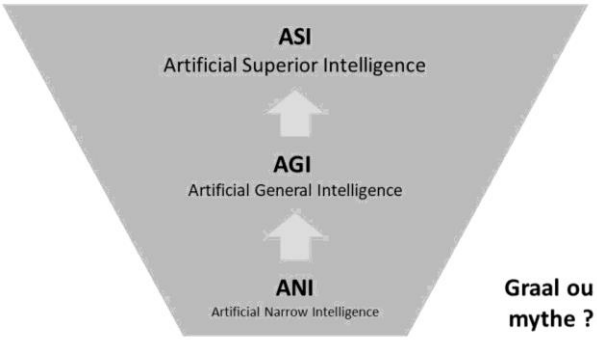
L'AGI nous ramène aux objectifs initiaux des pères de l'IA : intégrer l'intelligence humaine dans des machines, en pièces détachées ou de manière intégrée pour la rendre égale à celle de l'Homme dans ses différentes dimensions, notamment en termes de capacité de raisonnement et d'adaptation.

Cela reste un Graal très lointain mais divers chercheurs y travaillent. Les approches choisies sont variées avec le rapprochement de l'IA déductive et de l'IA inductive (ou symbolique vs connexionniste), avec la poursuite des progrès dans les neurosciences pour mieux comprendre le fonctionnement du cerveau et enfin, avec des approches complémentaires visant carrément à connecter le cerveau aux IA si ce n'est à le copier intégralement pour le faire tourner dans une machine. Un fantasme plus qu'autre chose compte-tenu de l'infinie complexité du cerveau.

Trois niveaux de l'IA

Au plus haut niveau conceptuel, on segmente habituellement l'IA en **IA forte** qui imiterait l'intelligence humaine avec capacité de raisonnement généraliste et **IA faible**, qui évoluerait de manière incrémentale à partir d'outils plus élémentaires, soit l'état de l'art actuel.

La distinction entre IA forte et IA faible se retrouve dans cette **classification** de la portée de l'IA avec trois niveaux d'IA : l'ANI, l'AGI et l'ASI, l'ANI correspondant à l'IA faible et les AGI et ASI aux IA fortes. C'est une classification très simpliste et bien trop anthropomorphique des évolutions futures de l'IA, très étroitement liées aux thèses de la singularité, qui anticipe ce moment fatidique où une machine atteindra puis dépassera l'intelligence humaine, ce qui n'a aucun sens quand on regarde les choses de près[148].



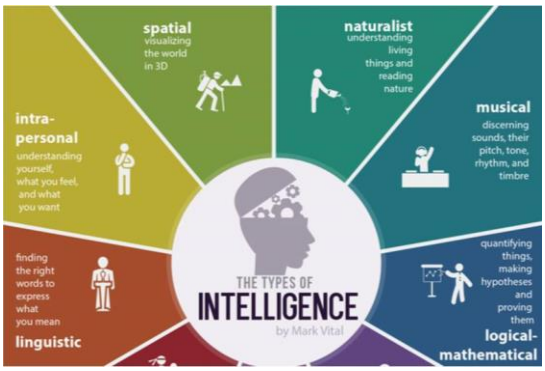
L'**Artificial Narrow Intelligence** (ANI) correspond à l'état de l'art actuel avec une IA capable de traiter des problèmes des domaines très précis. C'est ce que font aujourd'hui les IA exploitant aussi bien le machine learning que le deep learning ou les moteurs de règles. On peut y mettre en vrac les moteurs de recherche courants, la détection de fraudes bancaires, le credit rating de particuliers, la conduite automatique ou assistée, l'interprétation d'imagerie médicale, Apple SIRI, Amazon Alexa, Microsoft Cortana et Google Assistant.

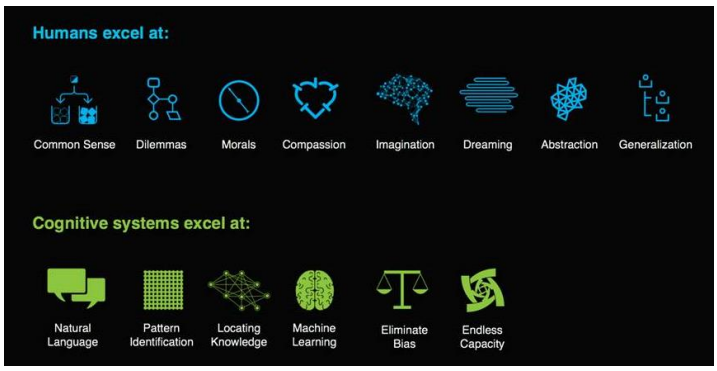
Si l'IA n'imite pour l'instant pas encore toutes les composantes de l'intelligence humaine, la force brute et l'usage d'éléments techniques dont l'homme ne dispose pas comme la vitesse de traitement et le stockage de gros volumes de données permettent déjà à la machine de dépasser l'homme dans tout un tas de domaines ! Et ce n'est pas nouveau ! Un tableur est déjà des millions de fois plus puissant qu'un humain doté des meilleures capacités de calcul mental ! La mémoire brute d'un Homme est très limitée.

[148] Si Ray Kurzweil est connu comme le grand promoteur de la singularité, son créateur est en fait Vernor Vinge dans The Coming Technological Singularity: How to Survive in the Post-Human Era en 1993 (12 pages). Il prévoyait l'échéance de la singularité pour 2023. Donc dans 5 ans. Et cette échéance correspondait au passage à la fin de l'humanité. Mais si l'Humanité court à sa perte, ce n'est pas forcément pas ce biais là mais bien plus à cause de questions environnementales.

On peut estimer que la mémoire verbale d'une personne moyenne est située aux alentours d'un Go de données ! Mémoire à laquelle il faudrait évidemment ajouter toute la mémoire sensorielle : visuelle, auditive et olfactive, qui est probablement très dense et maillée. Nos souvenirs sont en effet généralement multisensoriels et associatifs, une fonction qui est pour l'instant très difficile à reproduire dans des machines.

Dans le second étage, l'**Artificial General Intelligence** (AGI) correspondrait conceptuellement à un niveau d'intelligence équivalent à celui de l'Homme, avec un côté polyvalent, avec la capacité à raisonner, à modéliser les connaissances, analyser des données et résoudre des problèmes variés. L'AGI est en fait dans la continuité des travaux des pionniers de l'IA qui cherchaient à créer des systèmes d'IA capables de résoudre de manière générique toutes sortes de problèmes et en avaient même relancé l'idée au milieu des années 2000 dans l'initiative **HLAI** (Human Level AI).





On peut intégrer dans ce niveau un grand nombre des capacités humaines : l'usage du langage à la fois comme émetteur et récepteur, l'apprentissage par la lecture ou l'expérience, la mémoire et en particulier la mémoire associative, l'usage de la vue et les autres sens, le jugement et la prise de décisions, la résolution de problèmes multi-facettes, la création en général, notamment au niveau conceptuel mais aussi pour résoudre des problèmes physiques, la perception du monde et de soi-même, la capacité à réagir à l'imprévu dans un environnement complexe physique comme intellectuel ou encore la capacité d'anticipation[149]. Et à plus haut niveau, il faudrait intégrer la conscience, les sentiments, la sagesse et la connaissance de soi.

On pourrait y ajouter la capacité à ressentir des émotions que celles-ci soient personnelles (introspection) ou celle d'autres personnes (l'empathie), avoir des envies et des désirs et aussi savoir gérer ses pulsions et agir avec plus ou moins de rationalité. Ou alors, une machine n'aurait pas du tout besoin de tout cela pour résoudre des problèmes complexes inaccessibles à l'intelligence humaine. Cette liste est très longue ! Pour l'instant, on est encore très très loin de l'AGI, même si certaines de ces capacités notamment linguistiques et de raisonnement général sont en train de voir le jour.

Jusqu'à présent, les solutions d'IA fonctionnaient à un niveau de raisonnement relativement bas. Il reste à créer des machines capables de gérer le sens commun, une forme d'intelligence génétique capable à la fois de brasser le vaste univers des connaissances – au-delà de nos capacités – et d'y appliquer un raisonnement permettant d'identifier non pas des solutions mais des problèmes à résoudre. Il reste à apprendre aux solutions d'IA d'avoir envie de faire quelque chose. On ne sait pas non plus aider une solution d'IA à prendre du recul, à changer de mode de raisonnement dynamiquement, à mettre plusieurs informations en contexte, à trouver des patterns de ressemblance entre corpus d'idées d'univers différents permettant de résoudre des problèmes par analogie. On pourrait aussi développer des solutions d'IA capables de créer des théories et de les vérifier ensuite par l'expérimentation.

[149] On continue d'en découvrir tous les jours sur les principes biologiques de base de l'intelligence humaine, comme dans Brain Computation Is Organized via Power-of-Two-Based Permutation Logic publié fin 2016..

Pour ce qui est de l'ajout de ce qui fait de nous des êtres humains, comme la sensation de faim, de peur ou d'envie, d'empathie, de besoin de relations sociales, l'IA ne l'intègre pas. Ce n'est d'ailleurs pas nécessaire pour résoudre des problèmes courants auxquels s'attaquent les solutions à base d'IA.

Comme l'indique si bien **Yuval Noah Harari**, l'auteur du best-seller "Sapiens"[150], "*L'économie a besoin d'intelligence, pas de conscience*" ! On pourrait même ajouter, de rationnalité. Laissons donc une partie de notre intelligence voire une intelligence plus développée aux machines et conservons la conscience, les émotions et la créativité !

L'avènement éventuel d'une AGI dépend à la fois des progrès matériels et de notre compréhension toujours en devenir du fonctionnement du cerveau humain qui fait partie du vaste champ de la neurophysiologie, coiffant des domaines allant de la neurobiologie (pour les couches "basses") à la neuropsychologie (pour les couches "hautes").

Le fonctionnement du cerveau apparaît au gré des découvertes comme étant bien plus complexe et riche qu'imaginé. Les neurones seraient capables de stocker des informations analogiques et non pas binaires, ce qui en multiplierait la capacité de stockage de plusieurs ordres de grandeur par rapport à ce que l'on croyait jusqu'à il y a peu de temps.

On sait par contre que le cerveau est à la fois ultra-massivement parallèle avec ses trillions de synapses reliant les neurones entre elles mais très lent, avec une horloge tournant au grand maximum à 100 Hz. C'est aussi un engin très efficace du point de vue énergétique, ne consommant que 20W, par heure, soit l'équivalent d'un laptop équipé d'un processeur Intel Core i7 à pleine puissance.

Cette AGI est à mon sens une vue de l'esprit. Elle considère que l'IA ne se mesure que sur une dimension alors que, nous l'avons vu, elle s'évalue sur au moins deux douzaines de dimensions complémentaires. L'alignement de ces dimensions, la capacité de les rendre toutes supérieures à l'Homme dans la machine est une chimère. Nous aurons sans doute très longtemps des machines supérieures à l'Homme dans de nombreuses dimensions et inférieures dans d'autres, ce qui les rendra complémentaires à l'Homme. Des progrès seront certainement faits pour rendre les machines capables de raisonner par analogie pour résoudre des problèmes complexes mais elles n'intègreront pas toutes les composantes de l'intelligence humaine, notamment dans sa relation avec le monde physique.

Au troisième étage de cette fusée simpliste, l'**Artificial Super Intelligence** (ASI) serait une intelligence largement supérieure à l'Homme. Elle aurait même une intelligence infinie, tirant parti de la loi exponentielle de Moore. Une prévision qui se garde bien de rentrer dans les détails logistiques de l'application de cette loi ! L'échéance d'une telle AGI est située entre 2045 et 2062 selon les prévisions[151].

Ce serait la continuité logique et mécanique de l'étape précédente, liée à la puissance des machines qui se démultiplierait et se distribuerait plus facilement que celle d'un cerveau humain avec ses entrées-sorties et ses capacités de stockages et de traitement limitées et locales. Cette intelligence pourrait disposer de capteurs globaux : sur l'environnement, sur l'activité des gens, leurs déplacements, leurs loisirs, leurs états d'âme. Superintelligence va avec superinformation et super big data !

---

[150] Intervenant en juin 2016 dans la conférence USI organisée par Octo Technology à Paris (vidéo).

[151] En voici un exemple avec L'intelligence artificielle égalera l'intelligence humaine d'ici 2062 qui évoque les prévisions du chercheur australien Toby Walsh, auteur de 2062: The World that AI Made. Prévoir une AGI à 44 ans d'échéance semble évidemment bien trop précis. Reste à lire l'ouvrage, ce que je n'ai pas encore pu faire !

A vrai dire, une AGI serait d'emblée largement supérieure à l'Homme car elle accèderait facilement à tout le savoir humain déjà numérisé[152].

A ce niveau, l'intelligence de la machine dépasserait celle de l'homme dans tous les domaines y compris dans la créativité et même dans l'agilité sociale. Ce point de dépassement est une "singularité". Il est évoqué dans de nombreux ouvrages comme **The Singularity is Near** de Ray Kurzweil.

Pour de nombreux prospectivistes, l'ASI apparaîtrait très peu de temps après l'AGI, l'ordinateur faisant preuve d'une capacité à se reproduire lui-même, y compris à l'échelle matérielle. C'est évidemment une vue de l'esprit, tout du moins, lorsque l'on observe la manière dont fonctionnent les data-centers. Si ceux-ci étaient entièrement robotisés et alimentés en serveurs et systèmes de stockage par des camions autonomes eux-mêmes alimentés par des usines entièrement autonomes, pourquoi pas. D'où le besoin de préserver un minimum de contrôle humain dans cette chaîne de valeur.

Dans l'essai **The Singularity – A philosophical analysis,** le philosophe australien David J. Chalmers propose de tester d'abord l'ASI dans un environnement entièrement virtuel entièrement déconnecté du monde réel pour tester ses aptitudes. Si cela peut rassurer [153]!

Dans la plupart des prévisions sur l'avènement de l'ASI, il est fait état de la difficulté à la contrôler. Une majeure partie des prévisions envisagent qu'elle soit même néfaste pour l'homme malgré son origine humaine. Elles évoquent une course contre la montre entre startups et grandes entreprises pour être les premiers à créer cette ASI. Voir une course face à l'un des plus gros financeurs de l'IA : la DARPA.

Toutes ces conjectures semblent bien théoriques. Elles partent du principe qu'une ASI contrôlerait sans restriction toutes les ressources humaines. Elles s'appuient aussi sur la possibilité que toutes les sécurités informatiques d'origine humaine pourront être cassées par une ASI. C'est une vision dystopique et anthropomorphique du rôle des machines.

Les perpectives de l'AGI et de l'ASI sont surtout influencées par la science-fiction et ses grands auteurs : **Philippe K. Dick** (Blade Runner, Total Recall), **William Gibson** (Neuromancien, qui a inspiré la série de films **Matrix**), **Arthur C. Clarck** (2001 Odyssée de l'Espace), **Gene Roddenberry** (Star Trek), **Vernon Vinge** (le père spirituel du concept de la singularité dans « The coming technological singularity » en 1993, récupéré ensuite par Ray Kurzweil en 1999 dans « The age of spiritual machines »[154]) et bien entendu **Isaac Asimov** (AI, et ses fameuses trois règles de la robotique).

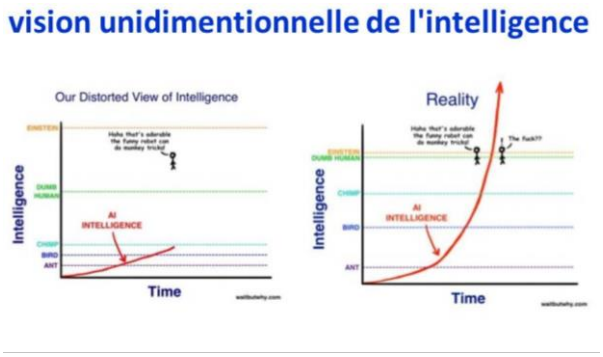
Les prédictions sur l'avènement de l'AGI sont souvent associées à un usage quelque peu abusif et prospectif de la loi de Moore. Or sa pérennité n'est pas garantie, quand bien même on pourrait mettre au point des ordinateurs quantiques dans les 40 ans qui viennent.

---

[152] Dans « The inevitable », publié en 2016, Kevin Kelly estime la production de contenu humaine depuis les sumériens à 310 millions de livres, 1,4 milliards d'articles et essais, 180 millions de chansons, 330 000 films de long métrage, 3,5 trillions d'images, un milliard d'heures de vidéo, télévision et courts métrages, et 60 trillions de pages web publiques. Et chaque année, le stock s'agrandirait avec 2 millions de livres, 16 000 films (dont nous ne voyons qu'à peine 1%), 8 millions de chansons et 30 milliards d'articles de blogs. Cela ne comprend pas les données brutes issues d'usages numériques (télécoms, réseaux sociaux, objets connectés). Cela représenterait 50 péta-octets de données. Avec les dernières technologies de stockage SSD, tout cela tiendrait dans un simple rack de data center. Voir le dernier SSD d'Intel. L'intégration de toute cette connaissance dans un réseau de neurones de deep learning se heurterait cependant à des limites techniques pas évidentes à surmonter. Mais avec un bon moteur de recherche, une AGI aurait toutefois une bonne capacité à exploiter cette base de connaissances en fonction des besoins.

[153] On peut aussi se rassurer avec ce très bon papier Ruper Goodwins paru en décembre 2015 dans Ars Technica UK : Demystifying artificial intelligence: No, the Singularity is not just around the corner.

[154] Voir Technoogical Singularity de Vernon Vinge (16 pages).



Heureusement, l'AGI a aussi eu son camp de détracteurs et dans toutes les décennies récentes (liste non exhaustive *ci-dessous*[155]). En France, le chercheur **Jean-Gabriel Ganacia**[156] auteur du « Mythe de la singularité » est assez remonté contre le mythe de l'AGI.

Dans **The Brain vs Deep Learning Part I: Computational Complexity — Or Why the Singularity Is Nowhere Near**, Tim Dettmers avance que la machine ne pourra pas dépasser le cerveau pendant le siècle en cours. Il démonte les prédictions de Ray Kurzweil[157].

Mais il n'est pas nécessaire de maîtriser le niveau d'abstraction le plus bas du cerveau pour en simuler les niveaux élevés, sans passer par un clonage. Comme il n'est pas nécessaire de maîtriser les bosons de Higgs pour faire de la chimie ou comprendre la manière dont l'ADN sert à fabriquer des protéines au sein des cellules !

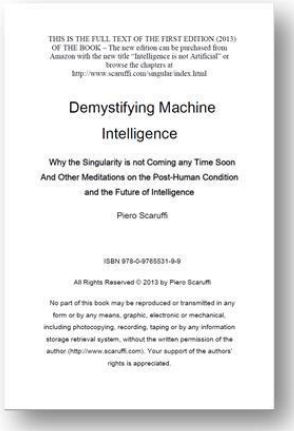
J. R. Lucas, Minds, Machines and Gödel, Philosophy XXXVI, 1961.

H.L. Dreyfus, "What Computers Still Can't Do", 1992.

R. Penrose, "The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics", 1989.

Nills Nilsson, "Human-Level Artificial Intelligence? Be Serious!", 2005.

D. Chalmers, Contemporary Philosophy of Mind: An Annotated Bibliography, Part 4: Philosophy of Artificial Intelligence.



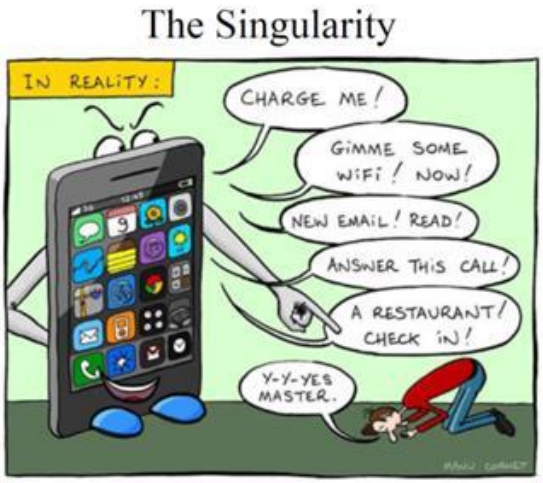
En tout cas, quoi qu'il arrive, l'intelligence d'une machine hyper-intelligente n'aura pas une intelligence similaire à celle de l'homme. Elle sera probablement plus froide, plus rationnelle, moins émotionnelle et plus globale dans sa portée et sa compréhension du monde.

[155] Voir aussi The Emperor of Strong AI Has No Clothes Limits to Artificial Intelligence de Adriana Braga et Robert Logan, 2017 (21 pages).

[156] Voir La Singularité, ça ne tient pas la route ! de Hubert Guillaud, juin 2017.

[157] J'avais moi-même émis des doutes sur les exponentielles qui sont la primitive des raisonnements de Kurzweil en avril 2015 dans trois articles sur La dérive des exponentielles. Voir aussi The Seven Deadly Sins of Predicting the Future of AI de Rodney Brooks, septembre 2017.

L'intelligence artificielle sera supérieure à celle de l'homme dans de nombreux domaines et pas dans d'autres, comme aujourd'hui. Elle sera simplement différente et complémentaire. Tout du moins, à une échéance raisonnable de quelques décennies.



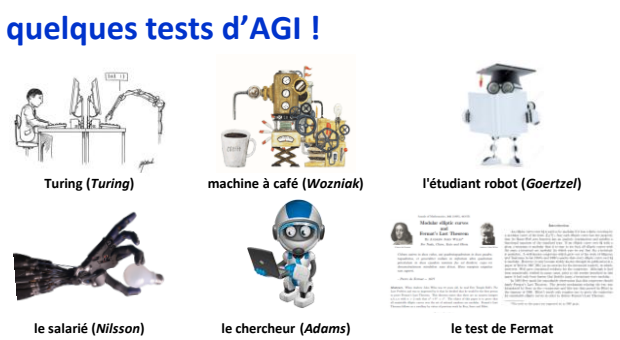
Enfin, il reste l'humour. Pour réussir le test de turing, il existe une solution très simple : rendre les gens moins intelligents ! C'est d'ailleurs l'impact qu'ont souvent les outils numériques, avec les différentes formes d'addiction qu'ils génèrent. Mais pas qu'eux. Le type de médias et contenus consommés ont aussi une forte influence. C'est la thèse ironique de **Piero Scaruffi** [158] et aussi celle de **Nicholas Carr**[159].

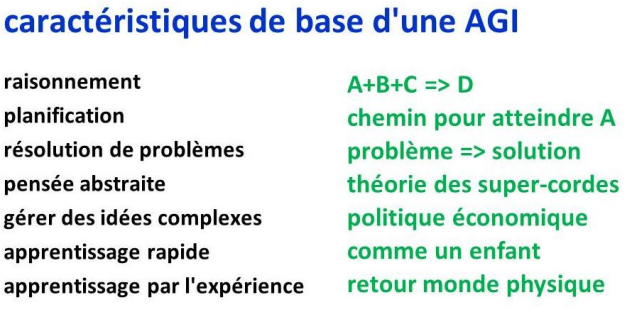
Pourquoi faire ?

Les futurologues qui s'avancent sur l'échéance de l'apparition d'une AGI oublient systématiquement de traiter une question clé : pourquoi faire ? Repartir des besoins et des problèmes permettrait peut-être de circonscrire le sujet et les recherches ! On aimerait bien qu'ils mettent sur la table quelques problèmes clés permettant d'améliorer le sort de la planète et de ses habitants. A la place, on a droit à des prévisions anxyogènes sur des AGI qui emprunteraient à l'Homme tout ce qu'il a de mauvais.

Quelques cas pratiques d'usage de l'AGI ont été définis par le passé mais ils sont très généraux, comme le **test de Turing** (agent conversationnel textuel que l'on ne peut pas distinguer d'un humain), le **test de la machine à café** de Steve Wozniak (un robot peut entrer dans un logement, trouver la machine à café, l'eau, le café et la tasse, faire le café et le servir), le **test de l'étudiant** robot (capable de suivre de cours et de passer avec succès les examens), celui du **chercheur** (capable de mener des travaux de recherche, de soumettre une thèse et d'obtenir son doctorat) et enfin, celui du **salarié** de tel ou tel métier, peut-être le cas le plus facile pour certains métiers répétitifs.

Les choses pourraient se compliquer pour une AGI si on lui demandait de démontrer le théorème d'incomplétude de **Godel** selon lequel "*dans n'importe quelle théorie récursivement axiomatisable, cohérente et capable de « formaliser l'arithmétique, on peut construire un énoncé arithmétique qui ne peut être ni prouvé ni réfuté dans cette théorie"* ou encore du dernier théorème de **Fermat** ($x^n + y^n = z^n$ est impossible pour un entier n supérieur à 2).



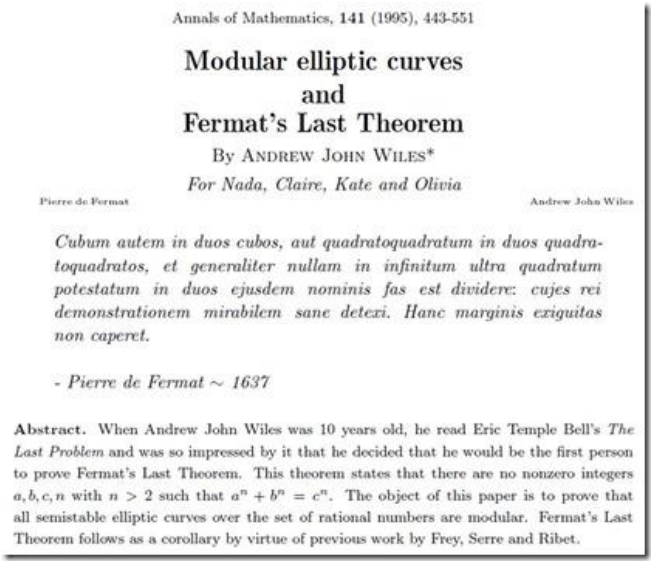


[158] Dans Artificial Intelligence and the Singularity, octobre 2014.

[159] Dans Is Google making us stupid, dans The Atlantic, juillet 2008.

Le théorème de Fermat me fascine. Il a été démontré par **Andrew Wiles** et après des années d'efforts de plusieurs mathématiciens. Sa démonstration publiée dans les annales de mathématiques en 1993 fait 109 pages et fait appel à de nombreux concepts incompréhensibles au commun des mortels, y compris pour votre serviteur passé par les classes préparatoires scientifiques au 20e siècle qui s'y perd à la troisième ligne de la démonstration.

Un **défi a été lancé en 2005** par un certain Jan Bergstra pour démontrer le théorème de Fermat avec un ordinateur et il reste toujours à relever. A vous de jouer si cela vous tente ! Le jour où une IA démontrera le théorème de Fermat, il y aura vraiment avoir de quoi être bluffé !



Annals of Mathematics, 141 (1995), 443-551

Modular elliptic curves and Fermat's Last Theorem

By Andrew John Wiles*

For Nada, Claire, Kate and Olivia

Pierre de Fermat — Andrew John Wiles

Cubum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos, et generaliter nullam in infinitum ultra quadratum potestatum in duos ejusdem nominis fas est dividere: cujes rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

- Pierre de Fermat ~ 1637

Abstract. When Andrew John Wiles was 10 years old, he read Eric Temple Bell's *The Last Problem* and was so impressed by it that he decided that he would be the first person to prove Fermat's Last Theorem. This theorem states that there are no nonzero integers a, b, c, n with $n > 2$ such that $a^n + b^n = c^n$. The object of this paper is to prove that all semistable elliptic curves over the set of rational numbers are modular. Fermat's Last Theorem follows as a corollary by virtue of previous work by Frey, Serre and Ribet.

Autre défi à relever pour une intelligence artificielle, résoudre les six défis mathématiques restants du **Clay Institute** lancés en 2000[160], avec $1M de récompense à la clé pour chacun d'entre eux. Cela comprend la démonstration des équations de Navier-Stokes dans la mécanique des fluides, la démonstration de P=NP ou P<>NP dans la théorie de la complexité[161] ou la démonstration de la conjecture de Poincaré !

Tant que l'on y est, on pourrait ajouter l'unification des théories de la physique des particules ou résoudre l'hypothèse de Riemann qui fait partie des 23 problèmes de Hilbert énoncés en 1900 et qui n'est toujours pas résolue.

Dans HolStep: a Machine Learning Dataset for Higher-Order Logic Theorem Proving, 2017, Cezary Kaliszyk de l'Université d'Innsbruck, François Chollet et Christian Szegedy de Google Research ont commencé à s'attaquer à ce problème en créant une base de connaissances ouverte classifiée exploitant les démonstrations de 11 400 théorèmes et plus de deux millions d'étapes intermédiaires de démonstrations. Reste à l'exploiter pour démontrer d'autres théorèmes et conjectures (théorèmes non encore démontrés) !

Bref, lorsque l'on parle d'AGI ou d'ASI, il vaut mieux non pas se comparer à l'Homme moyen, mais aux scientifiques et inventeurs les plus chevronnés de l'Histoire !

Comprendre le cerveau

Le concept même d'IA ne fait pas l'unanimité dans sa définition. Pour les puristes, un simple réseau de neurones ou un système de reconnaissance d'images ne relève pas à proprement parler de l'IA. Tout dépend de la définition que l'on se donne de l'IA, et notamment si la définition est anthropocentrée ou pas. C'est un peu comme la magie. Tant que l'on ne connaît pas le truc, c'est de la magie voire de l'art. Une fois qu'on le connaît, c'est une technique, souvent très simple, si ce n'est évidente.

L'intelligence humaine est un peu du même ressort quand on n'en connaît pas le fonctionnement exact. Elle préserve ce côté mystérieux et inimitable, presque immatériel, comme une âme qui n'aurait pas d'existence physique.

---

[160] Voir Clay Mathematics Institute – Millenium Problems. Un seul de ces défis à été relevé pendant les 18 ans qui nous séparent du lancement de cette initiative.

[161] J'évoque la question dans Comprendre l'informatique quantique – Complexité, juillet 2018.

L'intelligence humaine est magique tant qu'on a du mal à en expliquer le fonctionnement ou à connaître son origine exacte. C'est à la fois une affaire de neurosciences et de biologie (à bas niveau) et de sciences humaines (à haut niveau)[162].

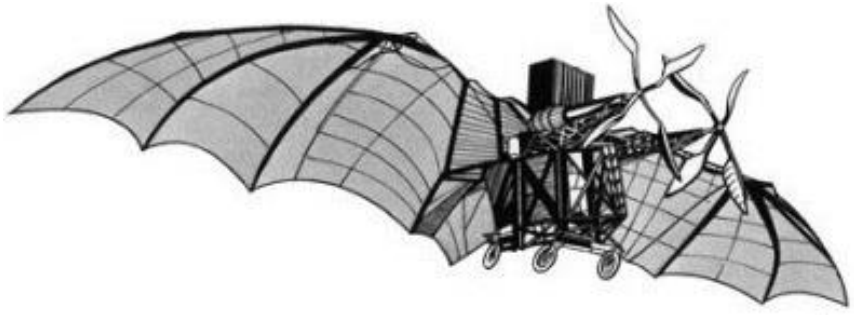
Au gré des découvertes en neurobiologie et en sciences cognitives, cette magie perd petit à petit de son lustre. L'homme n'est après tout qu'une machine biologique très sophistiquée issue de l'évolution. Certes, une machine complexe, une machine dont le fonctionnement dépend d'un très grand nombre de paramètres environnementaux et de l'accumulation d'expériences, mais une machine tout de même. C'est la première d'entre elles qui soit d'ailleurs capable d'en comprendre son fonctionnement interne ! Une plante ne connaît rien aux principes de la photosynthèse ou au cycle de Krebs qui décrit le matabolisme des glucides dans les cellules vivantes.

Doit-on absolument chercher à copier ou imiter le cerveau humain pour créer des solutions numériques ? Dans quel cas l'imitation est-elle utile et dans quels cas l'inspiration seulement nécessaire ? Doit-on chercher à créer des machines plus intelligentes que l'homme dans *toutes* ses dimensions ?

L'exemple de l'aviation couramment utilisée par Yann LeCun peut servir de bonne base de réflexion. L'avion s'inspire de l'oiseau mais ne l'imite pas pour autant. Les points communs sont d'avoir des ailes et d'utiliser la vitesse et la portance des ailes pour voler.

Le concept diverge alors rapidement : les avions n'ont pas d'ailes mobiles faites de plumes ! En lieu et place, leurs ailes sont généralement fixes et les moteurs sont à hélice ou sont des réacteurs. L'avion dépasse largement l'oiseau dans la vitesse (supersonique pour les avions militaires), la taille (B747, A380, Galaxy C5, Antonov 124), la capacité d'emport (qui se mesure en dizaines de tonnes), l'altitude (10 km pour un avion de ligne) et la résistance du froid (il y fait environ -50°C, ce qu'un organisme biologique développé peu difficilement supporter longtemps, même avec un bon plumage). Les avions sont par contre très inférieurs aux oiseaux côté efficacité énergétique et flexibilité, même si la densité énergétique de la graisse animale est voisine de celle du kérosène (37 vs 43 Méga Joules/Kg).

Le bio-mimétisme a été utile au début pour conceptualiser l'avion, que ce soit dans les schémas de Léonard de Vinci ou de l'avion de Clément Ader qui étaient très proches de l'oiseau. Si la motorisation d'un avion est très différente de celle des oiseaux qui battent de l'aile, les plumes se déployant au moment de l'atterrissage et du décollage sont cependant réapparues sous la forme des volets hypersustentateurs.



Ils sont été inventés par Boeing pour ses 707 lancés à la fin des les années 1950 (**description**) et dont la forme la plus élaborée a été intégrée aux Boeing 747 (*ci-dessus à droite*), dont les premiers vols ont eu lieu en 1969.

L'aigle est l'un des oiseaux les plus rapides au monde, atteignant 120 Km/h. Un avion de ligne classique atteint 1000 Km/h et il touche le sol, volets hypersustentateurs déployés, à environ 200 Km/h. Un A380 décolle en 2700 m et atterri sur 1500 m.

[162] Voir à ce sujet l'une des références des sciences cognitives, Stanislas Dehaene, avec ses cours au Collège de France. Il y décrit notamment l'origine du langage et des pensées.

Un aigle se pose en quelques secondes et presque n'importe où ! C'est la puissance et la capacité d'emport contre la flexibilité. Il faut se pencher du côté des drones de poche pour retrouver une part de la flexibilité des oiseaux mais leur autonomie est généralement bien plus limitée que celles des oiseaux, surtout les oiseaux migrateurs qui peuvent voler plusieurs heures d'affilée avant de se reposer au sol.

L'IA suit un chemin voisin dans le biomimétisme : certaines caractéristiques du cerveau des mammifères sont imitées dans les réseaux de neurones et le deep learning. Mais des différences fondamentales font diverger intelligence humaine et de la machine : à la fois ses entrées et sorties tout comme la structure de sa mémoire et du raisonnement. La machine se distingue pour l'instant par la capacité de stockage et d'analyse d'immenses volumes d'informations et par sa puissance de calcul brute.

L'homme dispose de capteurs sensoriels en quantité astronomique qu'aucun objet connecté n'égale à ce stade, ce qui, associés au cortex, lui procure une mémoire sensorielle qui accumule les souvenirs pendant toute son existence, provenant des entrées/sorties que sont les nerfs optiques, auditifs et olfactifs, ainsi que ceux qui gèrent le toucher, faits de millions de neurones irrigant en parallèle notre mémoire sensorielle[163]. C'est une force et une faiblesse. Nos émotions liées à cette mémoire sensorielle génèrent la peur de certains risques et des prises de décisions pouvant être irrationnelles. Ensuite, le niveau de complexité du cerveau dépasse l'entendement.

Il n'empêche que, par la force brute, l'IA dépasse déjà l'Homme dans tout un tas de domaines, notamment lorsqu'il faut "cruncher" de gros volumes de données qui nous échappent complètement. Quand elle a accès à de gros volumes de données comme dans l'oncologie ou en exploitant les données issues d'objets connectés, l'IA peut faire des merveilles.

Elle est d'ailleurs plutôt inopérante sans données. Elle ne sait pas encore quoi chercher ni prendre d'initiatives. Et les algorithmes sont encore très limités car les données de notre vie ne sont, heureusement, pas encore consolidées. Cela explique les limites de ces algorithmes de recommandation qui ne savent pas ce que j'ai déjà vu ou fait et ne sont pas prêts de le savoir. Ils ne peuvent donc pas faire de recommandation totalement pertinente. Le jour où toute notre vie sera suivie par des objets connectés depuis la naissance, il en sera peut-être autrement.

Qu'en est-il du raisonnement humain ? Celui-ci ne semble pas hors de portée des machines. On arrive petit à petit à le modéliser pour des tâches très spécialisées. Mais l'IA manque encore de souplesse et de capacité d'adaptation à une grande variété de situations. Bref, de jugeote !

Mais il n'est pas inconcevable d'arriver à fournir une intelligence générique à une machine. On y arrivera par tâtonnements, par intégration de briques algorithmiques et logicielles disparates, et pas seulement via la force brute de la machine[164].

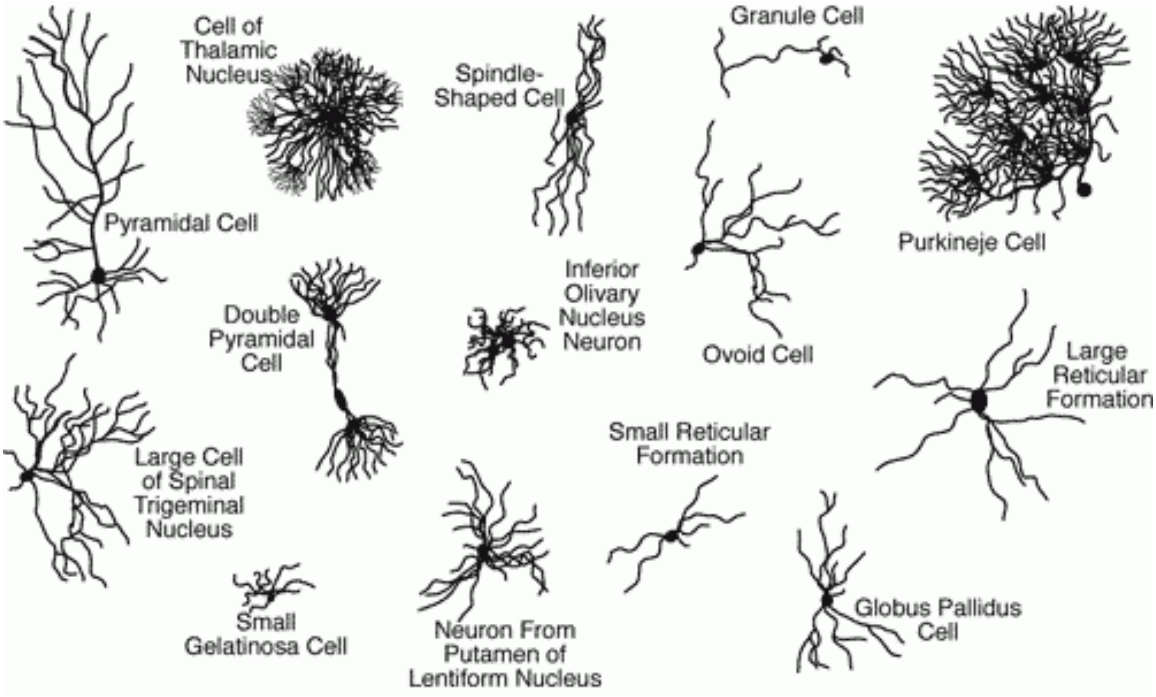
Je vais maintenant m'intéresser au fonctionnement du cerveau pour en évaluer la complexité et la difficulté à en modéliser le comportement dans de l'IA. Celui-ci contient plusieurs centaines de types de neurones différents[165], l'illustration *ci-dessous* n'en présentant que quelques grandes variantes. Le cervelet contient notamment ces étonnantes cellules de Purkineje, avec leur arbre de dendrites reliées avec jusqu'à 200 000 autres neurones, et qui contrôlent les mouvements appris.

---

[163] Brain Facts and Figures contient un imressionnant dimensionnement de l'ensemble du système nerveux humain. Quelques données lés : la moelle épinière qui relie le cerveau à l'ensemble du corps contient un milliard d'axones. Chaque rétine cintient 5 à 6 millions de cones qui captent la couleur et de 120 à 140 millions de bâtonnets qui captent l'intensité de la lumière et peuvent capter un seul photon. Le nerf optique contient 1,2 millions d'axones. Le cortex visual contient 538 millions de neurons. Dans l'oreille, la cochlée contient 15500 cellules cillées. Le nerf auditif contient 30 000 axones. Le cortex auditif contient 100 millions de neurones. Bref, côté sens, le cerveau est compliqué !

[164] Voir cette excellente présentation : Reverse-Engineering the Brain 2018 (192 slides) qui décrit l'état de l'art de la compréhension du fonctionnement du cerveau et essaye de comprendre l'origine de la plasticité cérébrale.

[165] Voir Neurons, Synapses, Action Potentials, and Neurotransmission.



(source du schéma ci-dessus sur quelques exemples de neurones cérébrales : **http://neuromorpho.org**)

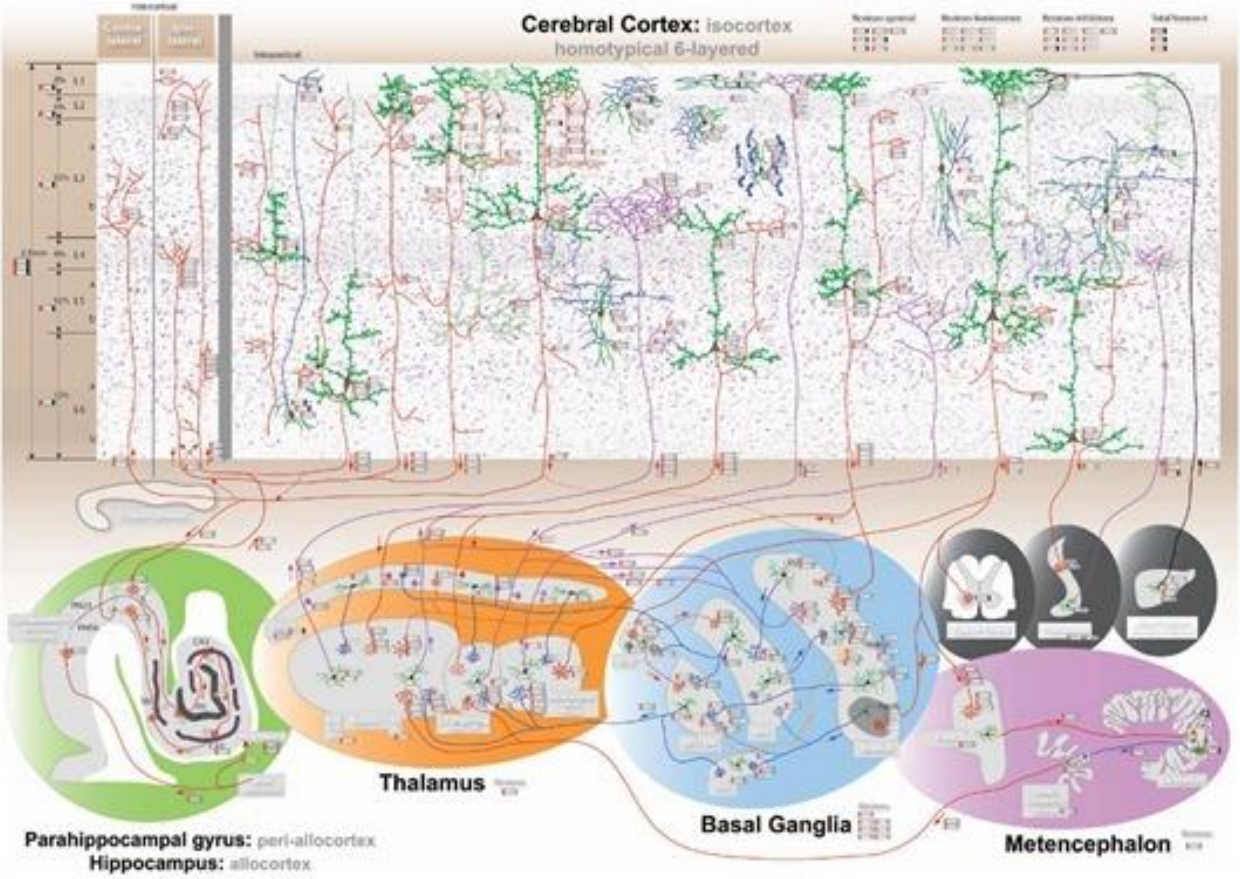
Cette complexité se situe aussi au niveau moléculaire avec de nombreuses protéines et hormones intervenant dans la transmission d'influx neuronaux[166]. Parmi les 25 000 gênes de nos cellules, 6000 sont spécifiques au fonctionnement du cerveau et leur expression varie d'un type de neurone à l'autre et en fonction de leur environnement !

Chaque gêne entraîne la création d'une grande variété de protéines du fait du principe de l'épissage qui fait se combiner dans des ordres variables les différentes composantes de l'ADN des gênes (les exons). C'est dire la richesse de la soupe de protéines qui gouverne le cerveau, dont l'actine qui structure la forme mouvante des neurones !

Le cerveau d'un fœtus comprendrait plus de mille milliards de neurones, qui meurent rapidement. On perd en fait des neurones dès sa naissance, comme si une matrice s'évidait pour prendre forme progressivement au gré des apprentissages. Le cerveau d'un enfant comprendrait plus de 100 milliards de neurones, et plus de 15 trillions de synapses et 150 milliards de dendrites.

Un cerveau adulte comprend environ 86 milliards de neurones dont 16 milliards dans le cortex et environ 56 milliards dans le cervelet. Ces neurones sont reliés entre eux par environ 600 trillions de synapses (liaisons neurones / neurones via les terminaisons multiples des axones qui sortent de neurones et se connectent aux dendrites proches du noyau d'autres neurones), et 300 milliards de dendrites (les structures des neurones sur lesquelles ne trouvent les synapses).

[166] La complexité protéinaire des neurons est décrite dans **Deep Molecular Diversity of Mammalian Synapses: Why It Matters and How to Measure It**.



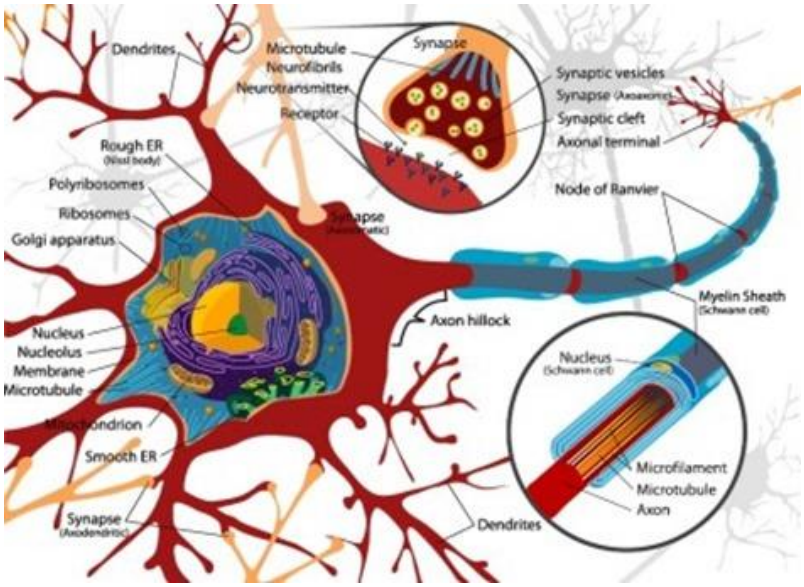
(source de l'illustration ci-dessus)

Les 20 Watts consommés par heure par le cerveau sont fournis sous forme d'hydrates de carbone (glucoses) via la circulation sanguine, ce qui en fait une "machine" très efficace côté consommation énergétique. Dans son développement à partir de la naissance, le cerveau perd des neurones mais gagne des liaisons entre elles, et ce, toute la vie, même si le processus se ralentit avec l'âge, même sans maladies neurodégénératives.

Un neurotransmetteur arrivant via une synapse peut déclencher une cascade de réactions en chaines dans le neurone cible qui va réguler l'expression de gênes et produire des protéines de régulation qui vont modifier le comportement des dendrites dans la réception des signaux issus des axones. Qui plus est les dendrites – les récepteurs dans les neurones – ont des formes et des comportements variables. Bref, nous avons un système de régulation des plus complexes qui n'a pas du tout été intégré dans les modèles Kurzweiliens !

Le cervelet cogère avec le cortex les automatismes appris comme la marche, la préhension, les sports, la conduite, le vélo, la danse ou la maîtrise des instruments de musique. Il intègre la perception sensorielle avec la commande motrice provenant du cortex. Un neurone du cervelet contient environ 25 000 synapses le reliant aux terminaisons d'axones d'autres neurones, avec l'exception des cellules de Purkinje qui ont 200 000 connexions.

Les neurones du cortex gèrent les sens et l'intelligence. Ils comprennent chacun de 5000 et 15 000 synapses.Le cerveau est aussi rempli de cellules gliales qui alimentent les neurones et en contrôlent le fonctionnement via la myéline qui entoure les axones et divers autres mécanismes de régulation. Il y en a au moins autant que de neurones dans le cerveau, ce qui ajoute un niveau de complexité de plus. Il faut ajouter le rôle de buffer de mémoire de l'hippocampe, le vidage de ce buffer pendant le sommeil ce qui rappelle que une bonne qualité et durée de sommeil permet d'entretenir sa mémoire.

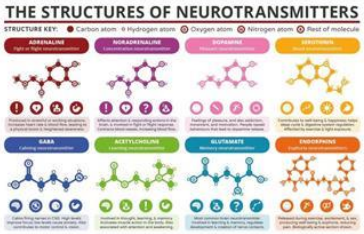


(**source** du schéma qui l'explique très bien)

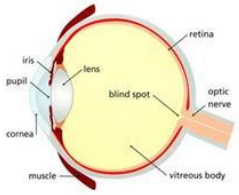
Enfin, via le système nerveux sympathique et parasympathique, le cerveau est relié au reste des organes, dont le système digestif ainsi qu'à tous les sens et notamment le toucher.

Le cerveau est imbattable dans sa densité, sa compacité et son parallélisme. Par contre, les ordinateurs nous dépassent dans leur capacité de stockage et de traitement de gros volumes de données. Si l'on aura bien longtemps du mal à scanner un cerveau au niveau des neurones, il n'en reste pas moins possible d'en comprendre le fonctionnement par tâtonnements.

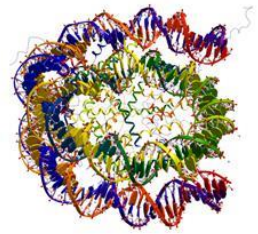
86 milliards de neurones
600 trillions de synapses



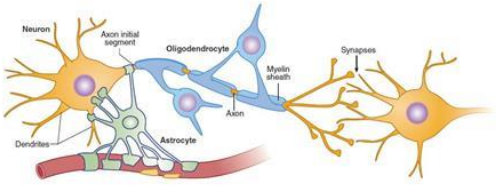
8 neurotransmetteurs différents



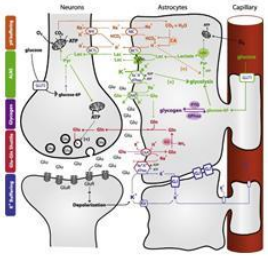
6,5 millions de cônes
90 millions de bâtonnets
1 million d'axones



6000 gênes spécifiques



88 milliards de cellules gliales dans le cerveau



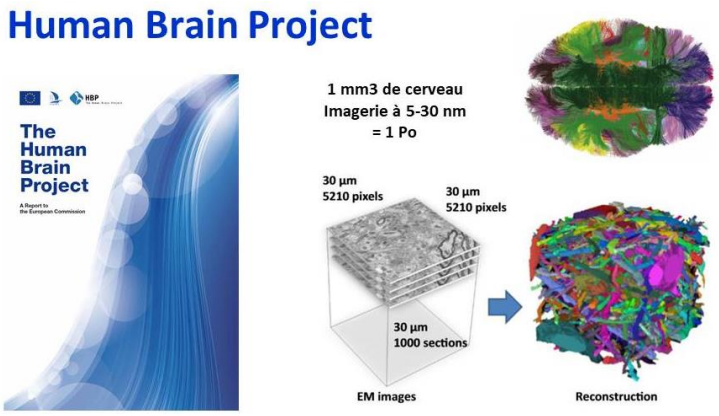
des interactions complexes

Les neurosciences continuent de progresser régulièrement de ce point de vue-là. On comprend petit à petit comment fonctionnent les différents niveaux d'abstraction dans le cerveau, même si les méthodes scientifiques de vérification associées restent assez empiriques, réalisées le plus souvent avec des souris.

Comprendre le cerveau en modélisant son fonctionnement reste cependant un objectif de nombreux chercheurs. L'idée n'est pas forcément de le copier, mais au moins de mieux connaître son fonctionnement pour découvrir des traitements de certaines pathologies neurodégénératives.

De nombreuses initiatives de recherche nationales et internationales ont été lancées dans ce sens. Elles sont issues d'Europe, des USA, mais aussi du Japon, d'Australie, d'Israël, de Corée et d'Inde.

Le projet européen **Human Brain Project** qui est un des trois « flagships » de l'Union Européenne avec les nanotechnologies et l'informatique quantique vise à simuler numériquement le fonctionnement d'un cerveau. Il a été lancé après la réponse à un appel d'offre par Henry Markram de l'EPFL de Lausanne. Ce chercheur est à l'origine du **Blue Brain Project** lancé en 2005, qui vise à créer un cerveau synthétique de mammifère.



Disposant d'un budget communautaire de 1Md€ étalé sur cinq ans, le Human Brain Project ambitionne de manière aussi large que possible d'améliorer la compréhension du fonctionnement du cerveau, avec en ligne de mire le traitement de pathologies neuro-cérébrales et la création d'avancées technologiques dans l'IA.

La cartographie à bas niveau du cerveau humain s'appuie sur le laboratoire **Neurospin** du CEA à Saclay qui est en train de mettre en place le système d'imagerie fonctionnelle par résonnance magnétique nucléaire le plus puissant du monde avec une résolution d'un dixième de millimètre[167].

A l'aide d'un supercalculateur Blue Gene d'IBM faisant tourner le logiciel de réseau de neurones de Michael Hines, le projet vise ensuite à simuler de manière aussi réaliste que possible des neurones[168] et un cerveau, en partant de celui de petits animaux comme les rats.

Le projet est critiqué ici et là[169]. Il fait penser un peu à Quaero par son aspect disséminé. Les laboratoires français ont récolté 78M€ de financement, notamment au CEA, tandis que ceux d'Allemagne et la Suisse se sont taillés la part du lion avec respectivement 266M€ et 176M€. On se demande qui fera l'intégration !

C'est plutôt un projet de big data qui s'éloigne du cerveau. En effet, les modèles de simulation ne s'appuient plus du tout sur la connaissance biologique actualisée que l'on a du fonctionnement des neurones dans le cerveau.

Les USA ne sont pas en reste avec la **BRAIN Initiative** annoncée par Barack Obama en 2013 et pilotée par le NIH (National Health Institute, équivalent US de l'INSERM français). Elle visait à mieux comprendre le fonctionnement du cerveau. L'objectif annoncé semble plus opérationnel que celui des européens : mieux comprendre les maladies d'Alzheimer et de Parkinson ainsi que divers troubles neuronaux.

---

[167] L'équipe de Neurospin a publié en août 2018 Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping, décrit dans Individual Brain Charting : une cartographie cérébrale à haute résolution des fonctions cognitives. L'état de l'art actuel est une cartographie du cerveau en IRM fonctionnelle avec une résolution spatiale de 1,5 mm.

[168] J'ai eu l'occasion de rentrer un peu plus en détails dans ce projet dans Ces startups qui veulent bidouiller le cerveau : les autres publié en juin 2017.

[169] Voir la critique très dure du HBP dans Dirty Rant About The Human Brain Project de Cathy O'Neil, 2015, qui note qu'il est bien trop tôt pour simuler le cerveau dans la mesure où l'on ne sait même pas décrire son fonctionnement correctement ainsi que Neuroscience: Where is the brain in the Human Brain Project? de Yves Frégnac et Gilles Laurent, 2014, qui s'interrogent sur la gouvernance du projet. J'ajouterai que la dernière conférence du HBP, le Human Brain Project Summit d'octobre 2018 n'est pas un bon gage de transparence : même son agenda est confidentiel !

Le budget annuel était de l'ordre de $100M, donc, in fine, du même ordre de grandeur que le Human Brain Projet. Parmi les projets, on trouve des initiatives en nano-technologies pour mesurer l'activité individuelle de cellules nerveuses, à commencer par celles des mouches drosophiles. Le plan initial est arrivé à son terme en 2018 et le NIH a lancé un appel à propositions pour le renouveler sur 5 ans.

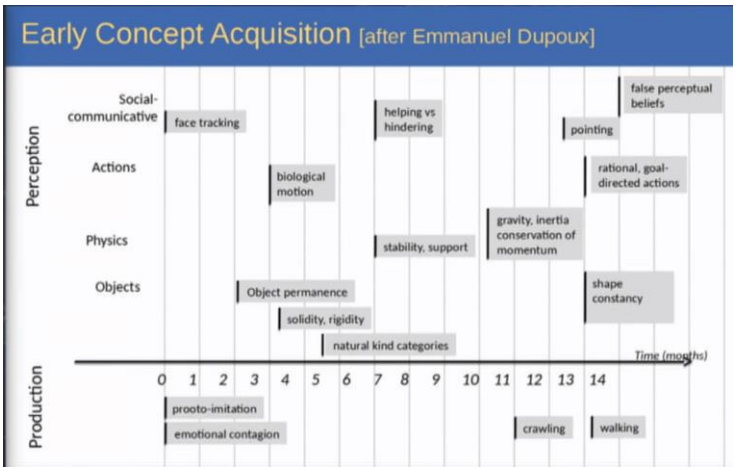
Dans le cadre de la BRAIN Intitiative, la **DARPA** a démarré en 2016 un projet visant à lire simultanément l'état de un million de neurones d'ici 2020 dans ce que l'on appelle la BCI (Brain Computer Interface)[170].

On peut aussi citer le **Human Connectome Project**, lancé en 2009, un autre projet américain, financé par le NIH comme la BRAIN Initiative, et qui vise à cartographier avec précision les différentes régions du cerveau.

De son côté, le projet **Allen Brain Atlas** planche sur la cartographie du cerveau de différentes espèces dont l'homme et la souris, au niveau de l'expression des gènes de ses différentes cellules nerveuses. La plateforme et les données associées sont ouvertes. Des chercheurs de l'Université de Berkeley ont même réussi à créer une cartographie précise de la sémantique du cortex[171].

Reste aussi, côté neurobiologie, à comprendre le processus d''apprentissage extraordinaire des enfants en bas âge et jusqu'à 20 ans. Comment le cerveau se câble-t-il pendant les phases d'apprentissage ? Comment séparer l'inné de l'acquis dans les processus d'apprentissage ?

Le slide *ci-contre* illustre les phases d'apprentissage d'un bébé et est issu de la conférence de Yann LeCun à l'USI à Paris en juin 2018 (vidéo)[172].



On dissèque les souris, mais bien évidemment pas les enfants en bas âge. Donc, on ne sait pas trop. Et l'IRM est insuffisante. Les chinois et les japonais planchent sur une voie intermédiaire en cartographiant le cerveau de singes qui sont plus proches de l'homme que les rongeurs.

Pour résumer, un bon nombre de recherches portent sur le fonctionnement du cerveau, avec une intersection avec les recherches en intelligence artificielle.

Hacker le cerveau

Une fois que l'on a compris le fonctionnement du cerveau, on peut envisager de le hacker. Soit pour le réparer, soit pour l'améliorer. Les réparations sont relatives au traitement de certaines pathologies neurodégénératives comme la maladie de Parkinson, qui est déjà en partie traitable par un hack utilisant une électrode implantée dans le cerveau. L'amélioration passerait par la création d'implants cérébraux permettant de connecter de manière bidirectionnelle le cerveau à des IA et au cloud.

Dans "The age of spiritual machines" de 1999, Ray Kurzweil prévoyait qu'en 2029, des nanorobots connecteront notre cortex cérébral au cloud. Or Kurzweil s'est régulièrement planté sur ses prévisions concernant les nano-technologies.

---

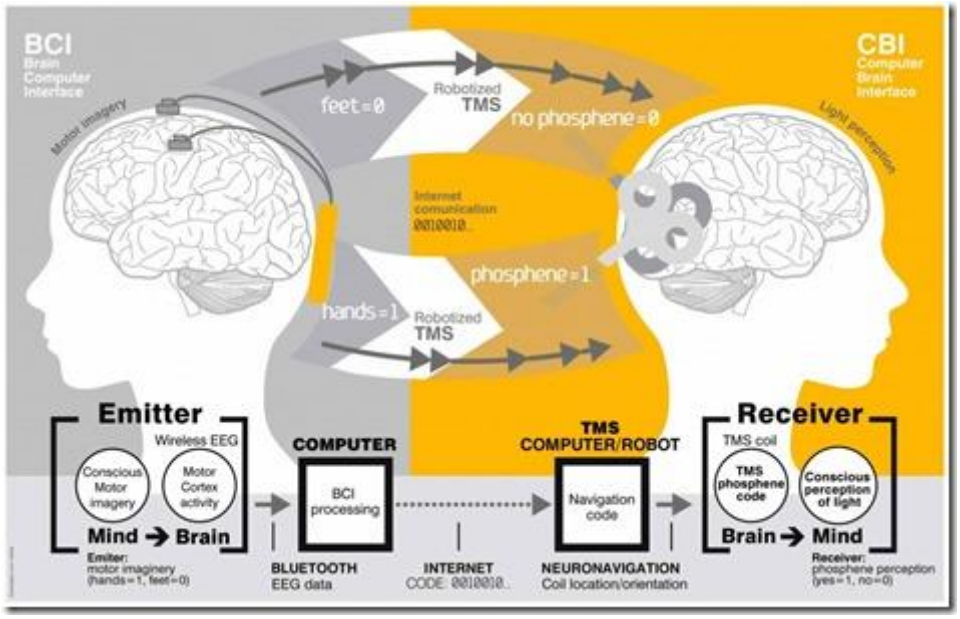
[170] Source : Government Seeks High-Fidelity "Brain-Computer" Interface, février 2016.

[171] Voir UC Berkeley team builds 'semantic atlas' of the human brain de Jessica Hall, 2016.

[172] Il précisait que les nombres de mois étaient mal alignés avec les fonctions apprises.

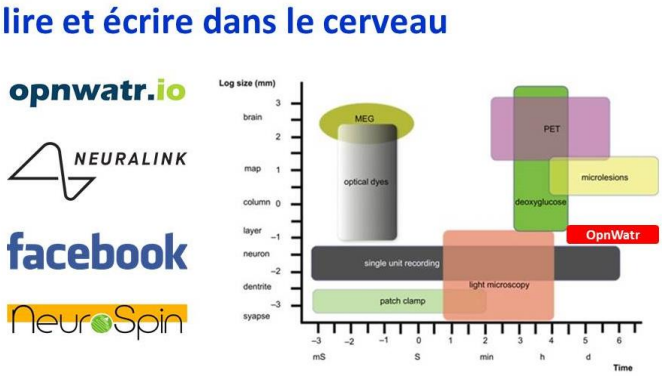
Il oubliait quelques petits détails pratiques de bon sens : comment créer des nano-robots pour chaque neurone du néocortex qui capteraient et influenceraient leur état, et le transmettraient à l'extérieur ? Il y a 16 milliards de neurones à cet endroit et 2000 fois plus de synapses au minimum. Ca fait beaucoup d'informations à collecter et à émettre à l'extérieur du cerveau ! Si on s'appuyait sur des émetteurs radio-fréquences, il serait intéressant d'évaluer l'énergie à consommer au sein du cerveau pour rendre cette communication possible, quand bien même celle-ci serait fournie par le glucose circulant dans le cerveau. On peut imaginer des nano-robots moléculaires. Mais aucun laboratoire n'a créé d'émetteurs d'ondes radio moléculaires. Ray Kurzweil se trompe au minimum sur le délai de la mise en oeuvre de sa prévision. Au mieux aura-t-on à cet horizon des prothèses qui relieront les entrées/sorties du cerveau à l'extérieur (vue, cortex moteur). Pas le néo-cortex directement.

Certains se lancent dans la connexion avec le cortex cérébral cognitif et visuel, et pas seulement moteur. Des **expériences de télépathie** sont possibles, en captant par EEG la pensée d'un mot d'une personne et en la transmettant à distance à une autre personne en lui présentant ce mot sous forme de flash visuel par le procédé TMS, de stimulation magnétique transcraniale. Mais on transmet un seul « bit » avec ce genre de technique. Un peu comme avec des casques de captation d'électroencéphalogrammes qui permettent de piloter par concentration un seul paramètre mécanique d'un jeu.



Si on peut déjà alimenter le cerveau au niveau de ses sens, comme de la vue, en interceptant le nerf optique et en simulant le fonctionnement de la rétine ou par la TMS, on ne sait pas l'alimenter en **idées** et **informations abstraites** car on ne sait pas encore vraiment comment et surtout où elles sont stockées. En tout cas pas encore car c'est l'ambition de startups américaines que d'y arriver un jour.

C'est le projet de **Neuralink**, une startup créé fin 2016 par Elon Musk et de **OpnWtr**[173] avec son bonnet utilisant des capteurs photos et des émetteurs infrarouges pour cartographier finement l'état des neurones et même, à terme, le modifier. **Facebook** essaye aussi de lire dans les pensées pour remplacer les claviers [174] ! Ces projets ont des niveaux d'ambition très différents. Elon Musk n'étant jamais à une exagération près, il prévoit que sa technologie de nano-électrodes permettra de lire et écrire dans les neurones à l'échelle individuelle.



C'est pour lui un moyen de relier l'Homme à l'IA pour qu'il puisse en tirer le meilleur parti et la contrôler. C'est négliger un point clé : la capacité d'une IA maléfique de hacker le cerveau du porteur d'électrodes.

Et si celles-ci ne peuvent être facilement enlevées ou débranchées, quel sera le recours de l'utilisateur en cas de problème ? Bref, c'est une solution potentiellement pire que le problème qu'elle est censée résoudre.

A ceci près que l'on n'a aucune idée de la manière dont est stockée la mémoire dans le cerveau et en particulier, que l'on ne sait pas relier sémantiquement chaque neurone du cortex à votre mémoire.

J'anticipe que les efforts de Neuralink n'aboutiront pas au point de permettre tout cela, mais qu'ils permettront tout de même de faire des progrès dans le traitement de certaines maladies neurodégénératives qui nécessitent de stimuler certaines parties précises du cerveau, surtout limbique (hypocampe, thalamus, hypothalamus). Ses premiers prototypes sont annoncés pour 2022.

C'est d'ailleurs ce qu'aurait réalisé un chercheur de l'**University of Southern California**, Dong Song, dans une expérience publiée fin 2017[175] avec sa solution pour améliorer la mémoire court terme de 15% à 30% (de quoi… ?[176]). Il s'agit d'un implant cérébral qui stimule électriquement l'hypocampe qui régit le fonctionnement de cette mémoire court terme. La stimulation électrique imite celle qui intervient dans les personnes saines. C'est censé servir aux personnes atteintes de la maladie d'Alzheimer ou de démence. Mais la stimulation est indiscriminée. Elle n'agit pas sur des neurones individuels ou sur des souvenirs précis.

C'est un peu ce qu'ambitionne de réaliser un certain Newton Howard d'Oxford avec l'implant neuronal Kiwi qu'il développe dans **Nitoo** (France). C'est une petite électrode qui capte des informations provenant des neurones ou agit dessus. La société est hébergée à Paris à l'Institut du Cerveau et de la Moelle Epinière et a été cofinancée par Bpifrance. La puce doit réduire les effets de maladies neurodégénératives comme celle d'Alzheimer ou de Parkinson.

Dans la lignée de Neuralink, nous avons aussi **Kernel** (2016, USA, $100M) qui souhaite créer une prothèse neuronale pour l'Homme[177] et plutôt dans la tendance « réparer l'Homme diminué ».

---

[173] J'ai publié en juin 2017 une étude détaillée des projets de Neuralink et OpenWatr dans une série de trois articles : http://www.oezratty.net/wordpress/2017/startups-bidouille-cerveau-neuralink/.

[174] Nanalyze a identifié 29 startups qui s'attaquent à l'interface cerveau-machine dans 29 Neurotech Companies Interfacing With Your Brain, octobre 2017.

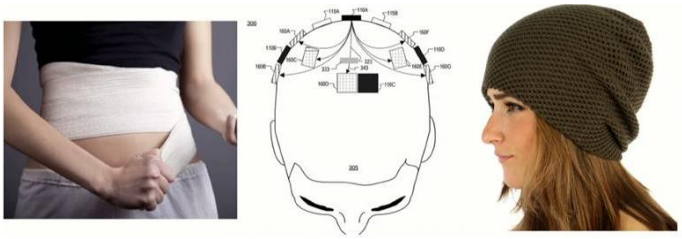
[175] Voir For the First Time Ever, Scientists Boosted Human Memory With a Brain Implant de Kristin Houser, novembre 2017.

[176] En fait, il semble que l'amélioration de performance concerne les personnes dont la mémoire court terme est défectueuse. Cela ne concerne pas les personnes en pleine forme.

[177] Voir Kernel's Quest to Enhance Human Intelligence de Bryan Johnson, octobre 2016.

Il faut évidement prendre avec des pincettes avec tous ces effets d'annonces. Ainsi, dans **Mashable**, une certaine Marine Benoit affirmait un peu rapidement en mars 2016 qu'une équipe avait mis au point "*un stimulateur capable d'alimenter directement le cerveau humain en informations*". A ceci près que l'étude en question, **Frontiers in Human Neuroscience** ne faisait état que d'un système qui modulait la capacité d'acquisition par stimulation, comme le projet de Dong Song ! Pour l'instant, on doit se contenter de lire dans le cerveau dans la dimension mécanique mais pas "écrire" dedans directement. On ne peut passer que par les "entrées/sorties", à savoir les nerfs qui véhiculent les sens, mais pas écrire directement dans la mémoire.

Chez **OpnWtr**, l'idée est d'utiliser leur bonnet pour réaliser des IRM fonctionnelles captant l'activité du cerveau à une résolution faible, de l'ordre du millimètre cube. C'est un outil de démocratisation des IRM qui aujourd'hui s'appuient sur des infrastructures lourdes en environnement hospitalier. Le projet d'OpnWtr est d'inverser à terme le processus et d'utiliser la technologie qui s'appuie sur des rayons infrarouge pour agir sur les neurones. Mais au vu de la résolution, l'impact sera très macro.



	MRI	Opnwatr.io
Depth	> 1 foot	> 1 foot
Resolution	1-3mm	Few mm -> sub mm
Cost	$2M + >$0.5M yearly	Consumer Electronics
Size	Entire Room	Wearable - Bandage or Ski-Hat

Enfin, **Facebook**, mène un projet pour capter les lettres et chiffres auxquels on pense avec un casque d'EEG (électro-encéphalogramme) et permettre une saisie plus rapide d'informations. C'est plausible d'un point de vue technique car une IRM fonctionnelle préalable permet d'identifier les zones associées dans le cortex. A ceci près qu'il n'est pas évident que cela permette une saisie de texte vraiment plus rapide qu'avec le doigt, un clavier ou la parole.

Ce projet est voisin de celui du MIT qui a pu capter via un EEG les mots auxquels on pensait[178]. Un réseau de neurones permet de faire la correspondance entre l'EEG capté et les mots auxquels on pense. Il doit probablement être entraîné pour chaque utilisateur.

Copier le cerveau

Après le hacking du cerveau, pourquoi ne pas copier directement son contenu dans une machine et nous faire vivre une vie virtuelle à la Matrix ? C'est des plus *border line* d'un point de vue technologique mais cela n'empêche pas certains d'y penser et d'y travailler.

Dans "The Singularity is Near[179]", Ray Kurzweil fantasmait sur la capacité à venir de transplanter un cerveau dans une machine et d'atteindre ainsi l'immortalité, incarnation ultime du solutionnisme technologique qui cherche à trouver une solution technologique à tous les problèmes ou fantasmes humains.

Le *dump* du contenu d'un cerveau dans un ordinateur fait cependant face à quelques obstacles technologiques de taille. Heureusement d'ailleurs[180] !

---

[178] Voir Computer system transcribes words users "speak silently, avril 2018 et la vidéo associée.

[179] L'ouvrage est librement téléchargeable dans The Singularity is near.

[180] Voir The empty brain de Robert Epstein, mai 2016, pour qui ce n'est même pas la peine d'essayer !

Quels sont-ils ? Tout d'abord, on ne sait pas encore précisément décrire le mode de stockage de l'information dans le cerveau. Se situe-t-il dans les neurones ou dans les synapses qui relient les neurones aux axones d'autres neurones ? Dans Memories may not live in neurons synapses paru dans Scientific American en 2015, il est fait état que l'information serait stockée dans les neurones et pas au niveau des synapses[181].

Ce stockage est-il du même ordre dans le cortex et dans le cervelet ? Qu'en est-il du cerveau limbique qui gère les émotions, le bonheur et la peur, en interagissant à la fois avec le cortex et avec les organes producteurs d'hormones ? On cherche encore !

L'information est proablement stockée sous forme de gradients chimiques et ioniques, probablement pas sous forme binaire ("on" ou "off") mais avec des niveaux intermédiaires. En langage informatique, on dirait que les neurones stockent peut-être des nombres entiers voire flottants au lieu de bits individuels. Il n'est pas exclu non plus que les neurones puissent stocker plusieurs informations à différents endroits (dendrites, synapses, axones, microtubules à l'intérieur des cellules) sans compter l'hypothèse « quantique » qui verrait les neurones stocker de grosses quantité d'informations dans les microtubules.

La communication entre deux neurones est chimique, via un potentiel d'ions calcium, sodium et potassium, et régulée par des hormones de régulation de la transmission nerveuse telles que l'acétylcholine, la dopamine, l'adrénaline ou des acides aminés comme le glutamate ou le GABA (acide γ-aminobutyrique) qui bloquent ou favorisent la transmission d'influx nerveux.

A cette complexité, il faut ajouter l'état des cellules gliales qui régulent l'ensemble et conditionnent notamment la performance des axones via la myéline qui l'entoure. La quantité de myéline autour des axones est variable d'un endroit à l'autre du cerveau et module à la fois l'intensité et la rapidité des transmissions nerveuses. Cela fait une complexité de plus dans le fonctionnement du cerveau !

Et si la mémoire n'était constituée que de règles et méthodes de rapprochement ? Et si le savoir était en fait encodé à la fois dans les neurones et dans les liaisons entre les neurones ? En tout cas, le cerveau est un gigantesque puzzle chimique qui se reconfigure en permanence.

Les neurones ne se reproduisent pas mais leurs connexions et la soupe biologique dans laquelle ils baignent évoluent sans cesse.

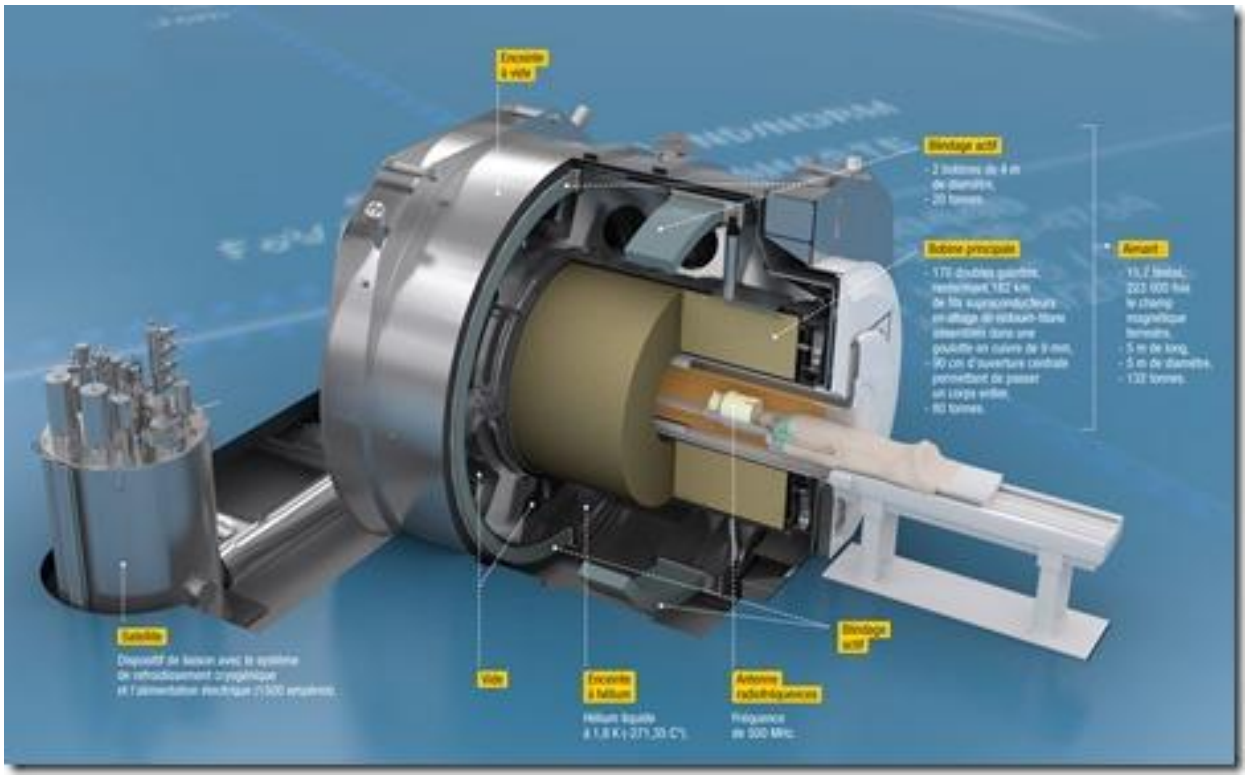
Comment détecter ces potentiels chimiques qui se trouvent à des trillions d'endroits dans le cerveau, soit au sein des neurones, soit dans les liaisons interneuronales ? Comment le faire avec un système d'analyse non destructif et non invasif ?

Il n'y a pas 36 solutions : il faut passer par des ondes électromagnétiques, et avec une précision de l'échelle du nanomètre. Aujourd'hui les scanners utilisent généralement trois technologiques : la tomodensitographie qui mesure la densité de la matière par rayons X, les PET scanners qui détectent des traceurs biologiques radioactifs par émission de photons et l'IRM qui détecte les corps mous par résonnance magnétique nucléaire, qui n'irradie pas le cerveau mais doit le plonger dans un bain magnétique intense. Ces scanners ont une résolution qui ne dépasse pas l'ordre du millimètre et elle ne progresse pas du tout en suivant une loi exponentielle de Moore !

Le dernier système en cours de mise en place dans le laboratoire **NeuroSpin** du CEA à Saclay, que nous avons déjà cité. Il s'agit du système franco-allemand **Iseult**, le scanner d'IRM corporel le plus puissant du monde, équipé d'un aimant record de 11,7 Telsas et 132 tonnes.

[181] Découverte confirmée par des chercheurs du MIT début 2016. Voir MIT discovers the location of memories: Individual neurons, mars 2012.

Son bobinage supraconducteur en niobium-titane refroidi par cryogénisation à l'hélium pèse 45 tonnes (*ci-dessous*, source). Il complètera l'IRM dotée d'un aimant de 7 tonnes qui est opérationnelle chez Neurospin depuis 2008. Plus l'aimant est puissant, plus on augmente la résolution de l'IRM[182].



Ce système va servir à générer des images 3D de plus haute résolution, descendant en-dessous du mm3 de l'IRM traditionnelle. Elle descendrait au niveau du dixième de mm (100 microns). Il est pour l'instant difficile d'aller en-deçà avec des techniques non invasives.

Iseult permettra d'identifier plusieurs types de molécules au-delà de l'eau, comme le glucose ou divers neurotransmetteurs, notamment via l'injection de marqueurs à base de molécules magnétisées. La mise en service prévue initialement pour 2018 est en retard de plusieurs années sur le calendrier initial. A terme, on pourra aller jusqu'à observer le fonctionnement des neurones à l'échelle individuelle.

Ce projet rappelle qu'une autre exponentielle est en jeu : plus on veut observer l'infiniment petit, plus l'instrument est grand et cher. Comme pour les accélérateurs de particule et le LHC du CERN pour la découverte du boson de Higgs. Plus on augmente la résolution de l'IRM fonctionnelle, plus il faut augmenter la fréquence de scan et la puissance de l'aimant, donc sa taille.

D'où l'intérêt de la solution légère et, en apparence, très élégante, de OpnWatr évoquée plus haut mais qui n'a pas encore fait ses preuves, notamment en termes de résolution spatiale.

Des capteurs d'électro-encéphalogrammes existent bien (EEG). Ils sont placés à la périphérie du cortex sur la tête et captent l'activité de grandes zones de contrôle psychomotrices du cerveau avec un faible niveau de précision.

[182] L'aimant a été conçu avec le concours du CEA-Irfu, l'Institut de Recherche sur les lois Fondamentales de l'Univers, qui a réutilisé ses acquis issus de la création des aimants supraconducteurs du Large Hadrons Collider du CERN de Genève. Il est fabriqué par Alstom-GE à Belfort, l'intégration du scanner étant réalisée par l'allemand Siemens, l'un des leaders mondiaux de l'IRM médicale. Y contribue également la société française Guerbet, spécialisée dans la production d'agents de contraste utilisés dans l'imagerie médicale.

C'est très "macro". La mémoire et le raisonnement fonctionnent au niveau du "pico". Qui plus est, si on sait cartographier approximativement les zones fonctionnelles du cerveau, on est bien incapable de capter le rôle de chaque neurone prise individuellement.

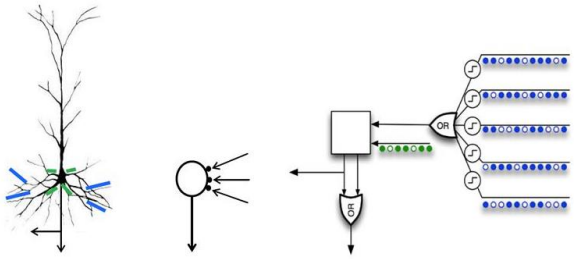
Pourra-t-on connaître avec précision la position et l'état de toutes les synapses dans l'ensemble du cerveau et à quels neurones elles appartiennent ? Pas évident ! Autre solution : cartographier le cortex pour identifier les patterns de pensée. Si on pense à un objet d'un tel type, cela rend peut-être actif des macro-zones distinctes du cerveau que l'on pourrait reconnaître. Mais cela demanderait un entraînement peronnalisé laborieux et de passer des heures allongé dans un scanner d'IRM.

L'autre solution, imaginée dans le cadre du Human Brain Project consisterait à congeler notre cerveau puis à le découper en fines tranches qui seraient ensuite scannées au microscrope électronique. Cela supposerait que le patient soit en phase terminal d'un cancer affectant d'autres organes que le cerveau. C'est ce que l'on appelle une technique destructive[183] !

Recherche en AGI

Le sujet de l'AGI est un thème de recherche actif. Il intègre de la recherche à haut niveau conceptuel pour bâtir des systèmes intégrant les mécanismes de l'intelligence et à un niveau pratique avec des startups qui ambitionnent de révolutionner le secteur, souvent en associant les notions de deep learning et divers systèmes de gestion des connaissances. Peu de chercheurs affichent ouvertement vouloir créer une AGI. Ils sont en général très spécialisés dans un domaine précis de la gestion des connaissances et du raisonnement.

Les tentatives affichées sont visibles chez quelques startups. Il y a d'abord **Numenta** (2005, USA), lancée par le créateur de Palm, Jeff Hawkins. La startup planche sur la simulation du cortex depuis plus d'une décennie avec ses réseaux de neurones HTM (Hiearchical Temporal Memory) mais qui n'a pas livré quoi que ce soit de bien concret à ce jour.



Elle fait du deep learning en cherchant à identifier des tendances temporelles dans les données pour faire des prévisions.

Leur solution Grok permet de détecter des anomalies dans des systèmes industriels et informatiques. Ils imitent le fonctionnement du cortex cérébral et de principes biologiques reprenant le principe de la mémoire par association et temporelle (Hierarchical Temporal Memory) théorisé par Jeff Hawkins en 2004 dans l'ouvrage **On Intelligence**, où il tente de décrire le fonctionnement du cerveau et la manière de l'émuler[184].

Les résaux de neurones à base de HTM utilisent des neurones plus sophistiqués que les réseaux de neurones habituels. Hawkins pense que le cerveau est principalement une machine prédictive qui n'est pas forcément dotée d'une capacité de calcul parallèle intensive mais plutôt d'une mémoire associative rapidement accessible. Il insiste sur l'importance du temps dans les mécanismes de rétropropagation mise en œuvre dans les réseaux neuronaux uniquement dans les phases d'apprentissage. Alors que le cerveau bénéficie d'une mise à jour sensorielle permanente.

---

[183] C'est ce que veut proposer la startup Nectome (USA, $120K). Voir A startup is pitching a mind-uploading service that is "100 percent fatal", mars 2018. Le proceed consiste à se faire embaumer le cerveau à basse temperature avec un liquide qui en préserve la structure. Ce sera proposé un de ces jours à des patients atteints de maladies incurables. Juste au moment de leur décès. Tant que ce n'est pas un cancer du cerveau, cela peut donner l'espoir de revivre un jour à la Matrix dans un ordinateur. Mais les chances sont plutôt réduite que ces clients soient un jour véritablement satisfaits.

[184] L'ouvrage est téléchargeable gratuitement ici. Mais le lien n'est pas sécurisé.

Les thèses de Hawkins sont intéressantes et constituaient un pot-pourri des connaissances en neurosciences il y a plus de 10 ans maintenant. Elles sont évidemment considérées comme un peu simplistes[185].

J'ajouterai à ces critiques que Hawkins oublie négligemment le rôle du cervelet et du cerveau limbique dans les apprentissages et le prédictif. Le cervelet contient plus de neurones que le cortex et il gère une bonne part des automatismes et mécanismes prédictifs, notamment moteurs.

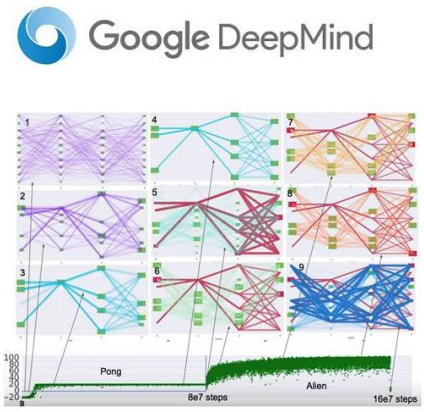
Numenta propose aussi NuPIC (Numenta Platform for Intelligent Computing) sous la forme d'un projet open source. Cette société est très intéressante dans le lot car elle utilise une approche technique plutôt originale qui dépasse les classiques réseaux neuronaux.

En octobre 2018, Numenta faisait reparler d'elle après un long silence en annonçant une grande découverte sur le fonctionnement du cerveau sous l'appellation « The Thousand Brains Theory of Intelligence ». En gros, ils pensent avoir découvert comment fonctionnait la mémoire spatiale dans le néocortex au niveau des colonnes corticales[186]. Cela reposerait sur les *grid cells* qui capturent la position de notre corps par rapport aux objets qui nous entourent. Les grid cells ont été découvertes en 2005. A ce stade, ces travaux permettraient d'avancer dans la conception d'AGI en s'appuyant sur du biomimétisme, et notamment pour la conception de robots capables de bien se mouvoir dans l'espace de manière autonome.

Il y a aussi le projet **MICrONS** financé par l'IARPA, la DARPA du renseignement, et mené par George Church qui voulait carrément émuler le cerveau. Le projet a été mené entre 2014 et 2015 et ne semble pas avoir eu de suite[187].

La filiale de Google **DeepMind** fait des avancées régulières dans le chemin torteux qui mène à l'AGI en s'appuyant sur la notion d'apprentissage neuro-symbolique, qui fusionne les réseaux de neurone du deep learning et la gestion des connaissances[188].

DeepMind présentait **PathNet** en 2016, un réseau d'agents intégrant des réseaux de neurones capable d'identifier la meilleure combinaison de réseaux de neurones pour atteindre un objectif donné. En 2017, ils publaient A simple neural network module for relational reasoning, 2017 (16 pages) dédié à la mémoire associative et en 2018, Measuring abstract reasoning in neural networks[189] (10 pages) pour modéliser du raisonnement abstrait de réponses à des tests élémentaires de QI.



Pathnet : architecture modulaire de deep learning.

"PathNet: Evolution Channels Gradient Descent in Super Neural Networks" (Fernando et. Al, 2017)

réseau de réseaux de neurones qui teste de nombreuses combinatoires de réseaux de neurones pour trouver le meilleur chemin vers la solution.

sorte de "méta deep learning"

[185] Voir ces critiques : Book Review: On Intelligence by Jeff Hawkins de Jeff Kramer, 2013, On Biological and Digital Intelligence de Ben Goertzel, et Is the model for general AI from On Intelligence by Jeff Hawkins reasonable and is it possible to use it practically?, 2014.

[186] Voir A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex, octobre 2018 (15 pages) ainsi que Locations in the Neocortex A Theory of Sensorimotor Object Recognition Using Cortical Grid Cells, 2018 (16 pages). Et l'explication simplifiée dans Numenta Introduces Breakthrough Theory for Intelligence and Cortical Computation, octobre 2018.

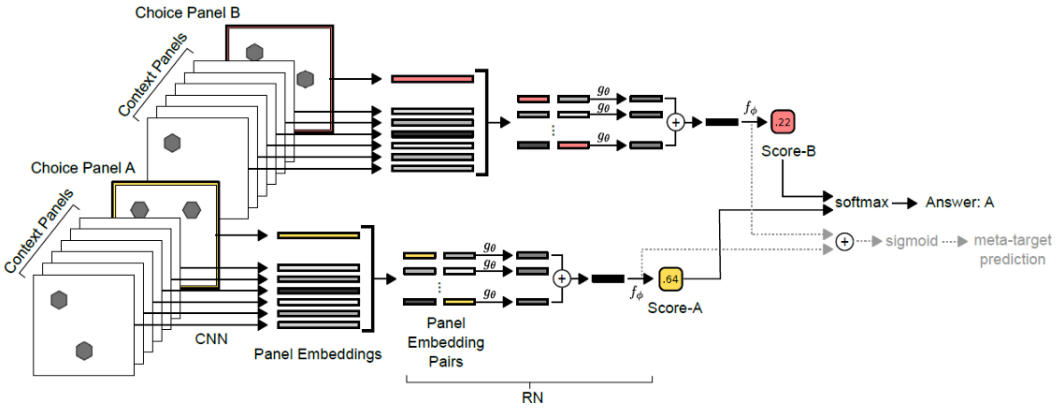
[187] Voir Machine Intelligence from Cortical Networks (MICrONS) ainsi que Hybrid bio-opto-electronics for AI avec une vidéo de Georges Church présentant le projet.

[188] Voir Neurosymbolic Computation Thinking beyond Deep Learning, 2017 (43 slides), Neural-Symbolic Learning and Reasoning A Survey and Interpretation, 2017 (58 pages) et Toward Neural Symbolic Processing de Hang Li, 2017 (36 slides). Côté outils pour le neuro symbolique, il y a notamment la bibliothèque CILP++ lancée en 2015.

[189] Vulgarisé dans What's New In Deep Learning Research: An IQ Test Proves that Neural Networks are Capable of Abstract Reasoning de Jesus Rodriguez, juillet 2018.

Le tout, dans un dispositif comparant plusieurs types de réseaux de neurones (un réseau convolutionnel, un ResNet, un réseau à mémoire LSTM, et un réseau nouveau relationnel WReN, plus performant, *ci-dessous*).

A vrai dire, on est toujours très loin d'une faculté de raisonnement généraliste puisque ce réseau de neurone est dédié à la résolution d'un seul problème particulier et relativement simpliste.

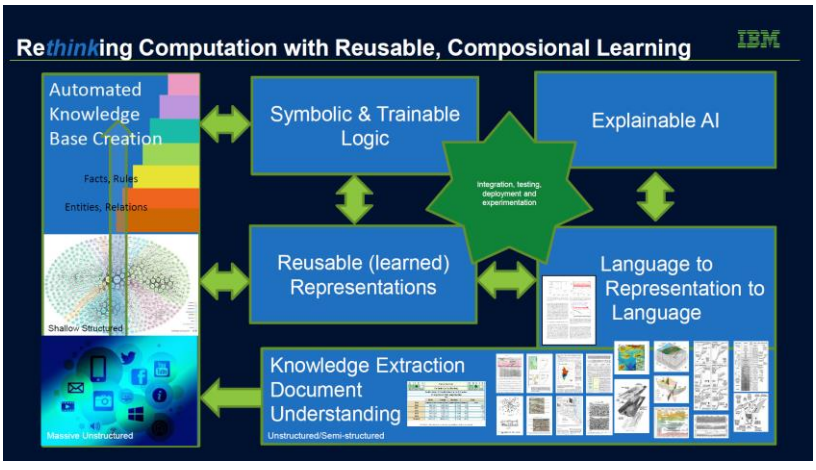


Nnaisense (2014, Suisse) a été créée par Jürgen Schmidhuber, le chercheur à l'origine des réseaux à mémoire LSTM utilisés dans le traitement du langage et la traduction. Ils ambitionnent de créer une AGI et dans un premier temps visent la finance, les industries lourdes et les véhicules autonomes. Comme ils travaillent en mode projet, ils ont déjà un chiffre d'affaire conséquent de $11M (2016). Mais ils sont encore très loins de l'ombre d'une AGI.

Des startups cherchent plus prosaïquement à faire coexister les approches symboliques à l'ancienne et connexionnistes. C'est par exemple le cas du Français **Nexyad** qui développe des solutions pour les véhicules autonomes, dont **SafetyNex** qui analyse les risques en temps réel en associant deep learning – ne serait-ce que pour exploiter les contenus issus des capteurs – et du raisonnement ainsi que de la logique floue.

Cela permet de créer une IA explicable, ce qui est critique pour la conduite de véhicules autonomes. Mais on n'est pas encore véritablement dans le champ de l'AGI car ces systèmes sont dédiés à des tâches très spécialisées.

Dans AI for Complex Situations: Beyond Uniform Problem Solving, 2017 (31 slides), Michael Witbrock, qui traaille dans les équipes d'**IBM Watson** et ancien de CyCorp, décrit la démarche de fusion de l'IA symbolique et de l'IA connexionniste pour la gestion des connaissances. Le schéma *ci-contre* l'illustre avec en bas, l'extraction des connaissances par l'analyse du langage dans des documents.



Cela permet d'alimenter une base de connaissances (à gauche) elle-même exploitée par une IA symbolique (au milieu en haut). Au passage, ce genre d'IA est explicable.

La création d'AGI doit surmonter de nombreuses difficultés, la première étant le manque de formalisme et d'abstraction qui est utilisé pour décrire et décomposer les différentes composantes de l'intelligence humaine.

C'est la thèse de Mike Loukides dans l'excellent Why we need to think differently about AI août 2018. Il souligne aussi le manque de rigueur scientifique de l'IA, qui relève le plus souvent de bricolage et d'une approche expérimentale pas toujours facilement reproductible manquant de fondements scientifiques, par exemple dans les mathématiques ou la philosophie.

Lorsqu'un domaine scientifique est bloqué, cela veut peut-être dire qu'il ne s'attaque pas aux bons problèmes. C'est peut-être le cas des approches qui cherchent à fusionner les moteurs de règles et le deep learning, ou les méthodes symboliques et connexionnistes.

L'autre écueil de l'IA est que la mesure de son progrès est au mieux approximative. Comme la mesure est une abstraction d'une abstraction, si la première abstraction est erronée, la méta-abstraction le sera tout autant.

L'autre point lié à la créativité dans l'intelligence humaine est que celle-ci consiste souvent à casser des règles existantes voir de changer de mode d'abstraction. Or les systèmes d'IA créatives à base de réseaux génératifs sont basés sur la combinaison de modèles et données existants. Ce qui génère un fonctionnement en boucle, comme lorsque des générateurs de musique esssayent d'imiter le style des Beatles. Ces modèles n'auraient pas pu créer la musique des Beatles avec celle qui les précédait !

Dans le très long How should we evaluate progress in AI?, 2018, **David Chapman** décrit la nature multi-dimensionnelle de l'IA qui associe des sciences, de l'ingéniering, des mathématiques, de la philosophie, du design et de la communication. Cela rend difficile la convergence autour de méthodes et approches communes.

L'IA est une discipline hybride et très expérimentale. Parfois, les méthodes formalisent trop vite les choses au lieu de rester dans le flou. Les travaux courants dans le deep learning génèrent des résultats Z avec des méthodes X et des données Y, parfois avec des progrès W par rapport à l'état de l'art, mais que l'on sache pourquoi ni que l'on puisse généraliser le résultat. C'est par exemple le cas des Capsule Networks qui fonctionnent pour l'instant avec des lettres mais pas encore avec des images, sans que le passage de l'un à l'autre soit explicité.

David Chapman décrit le risque de voir la recherche en IA affectée des mêmes syndromes qui ont touché les sciences de la psychologie il y une décennie, dont les travaux étaient entâchés de nombreuses fraudes et expériences difficiles à reproduire, ce qu'il appelle la crise de la réplication, aussi courante dans les biotechs[190].

Des conclusions erronées sont générées par les chercheurs lorsqu'ils communiquent sur leurs travaux en exagérant leurs effets, en interprétant les résultats de manière approximative, en ne relatant pas les essais infructeux de leurs recherches, en ne réalisant pas leurs expériences à une échelle suffisante, notamment d'un point de vue statistique[191].

Dans Troubling Trends in Machine Learning Scholarship de juillet 2018, Zachary Lipton propose quelques recettes pour rendre les publications scientifiques de l'IA plus propres en évitant de mélanger explications et spéculation, en décrivant l'origine des améliorations de performance obtenues, en évitant de noyer le lecteur dans un jargon mathématique cryptique et en étant bien précis sur les termes employés. Histoire d'éviter ce que l'on peut lire dans In the Future, There Will Be No Limit to What AI Can Accomplish in Science de Peter Rejce, mars 2018, qui survend effrontément les capacités existantes et à venir de l'IA. Heureusement, d'autres sont plus sages sur les prédictions, comme la chercheuse Melanie Mitchell dans Opinion | Artificial Intelligence Hits the Barrier of Meaning, novembre 2018, qui rappelle que les prévisions sur l'émergence d'AGI se sont régulièrement heurtées au mur de la difficulté de la réaliser.

---

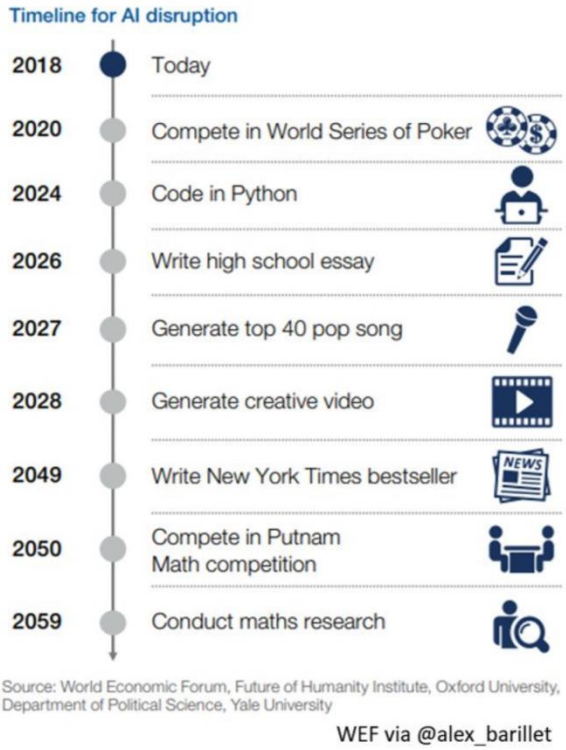
[190] Voir We need to improve the accuracy of AI accuracy discussions de Danny Crichton, mars 2018

[191] Exemple de biais incroyable avec cette étude New Use of A.I. Accurately Detected Cancer 86 Percent of the Time qui relate une IA qui détecte avec exactitude 86% de cancers colorectaux sur un échantillon de patients dont on savait déjà qu'ils étaient atteints de cette pathologie, créant un biais statistique énorme.

Il se trouve que l'IA a des limites, ne serait-ce que dans la capacité à résoudre des problèmes trop complexes (au-delà de NP-Difficile dans les classes de complexité) et ceux qui sont indécidables ! Selon le théorème d'incomplétude de Gödel, un système logique ne peut être à la fois complet et cohérent.

Il contient donc souvent des propositions indécidables. Une AGI devrait donc assister la prise de décisions faiblement formalisées et non simulables. Elle ne pourra pas décider à notre place face à l'incertitude surtout lorsqu'il s'agit d'innover [192]! Il en va de même pour les prédictions farfelues comme celles de la timeline *ci-contre*.

Il faut donc nous résigner à accepter le monde dans la complexité et la grande variété des paramètres qui le gouvernent. On ne peut pas tout prédire !



Dans Why we need to think differently about AI, Mike Loukides décrivait aussi les errements de la période GOFAI de l'IA, celle des raisonnements à base de moteurs de règles. Les critiques sur le deep learning d'aujourd'hui ne manquent pas non plus. On en trouve dans Deep Learning: Diminishing Returns? de Bernard Murphy, juillet 2018, dans Deep Learning: A Critical Appraisal de Gary Marcus, 2017 (27 pages) et dans son post dans Medium In defense of skepticism about deep learning, janvier 2018.

Selon Gary Marcus, les ConvNet (réseaux de neurones convolutifs) requierent trop d'informations et de calculs. Ils appliquent une forte brute trop brute et consommatrice de ressources machine. Il voit comme défis au deep learning le besoin de mieux gérer les informations hiérarchiques surtout dans le langage, le besoin de mieux différencier corrélations et causalités dans les analyses, comme l'Homme et de « désiloiser » le deep learning vis-à-vis de l'IA symbolique.

Comme Yann LeCun[193], il pense qu'il faut surtout développer les formes d'apprentissage non supervisé et les modèles hybrides symboliques/neuronaux qui observent le monde et en créent des modèles prédictifs à base de sens commun.

On commence à entendre ce message d'apprentissage non supervisé magique, par exemple chez Intel ou même avec Bruno Maisonnier d'AnotherBrain qui font la promotion de chipsets qui réduisent les tâches d'apprentissage sans que l'on sache d'ailleurs très bien à quoi ils servent.

Pour Yann LeCun, le problème du deep learning est que sa mémoire est limitée. D'où l'approche de Jeff Hawkins avec le HTM de Numenta et la mémoire temporale, même si cela n'a pour l'instant pas encore aboutit. Il pense aussi que pour faire du raisonnement, il faudrait générer des résultats multiples et pas uniques avec des modèles basés sur les contraintes énergétiques, ce qui rappelle le concept des modèles génétiques. Bref, il faudrait combiner les approches neuronales et les approches par graphes.

[192] Je m'inspire ici en le paraphrasant de Jean Staune dans « Les clés du futur », page 594.

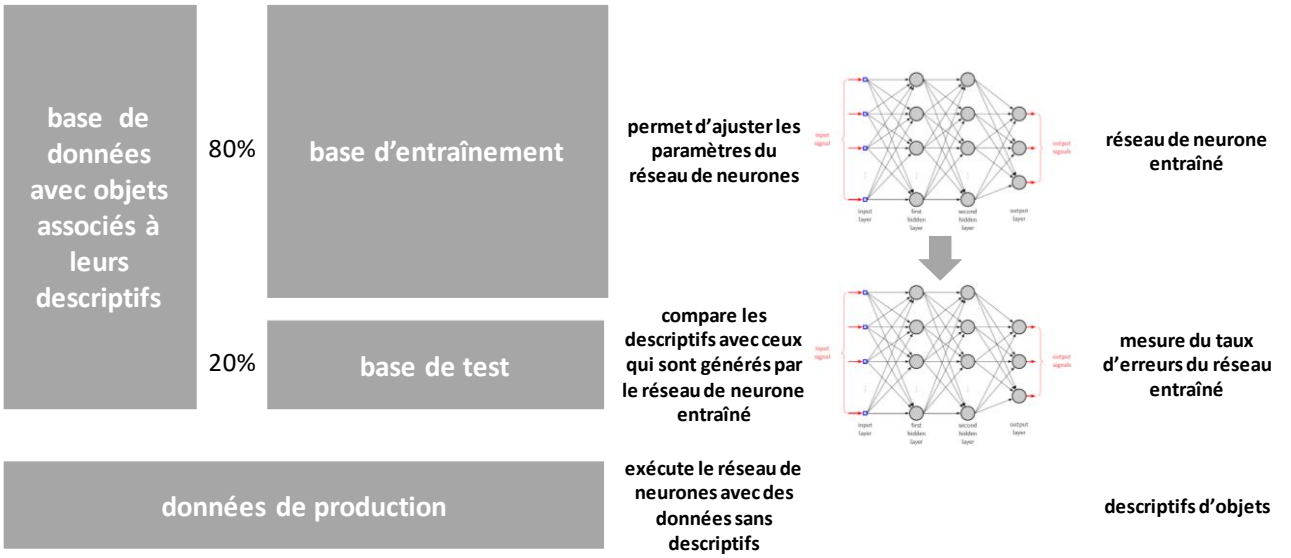
[193] Voir aussi Hinton, LeCun, Bengio : la « conspiration » du deep learning, de Benoît Georges dans Les Echos, juillet 2018.

Données de l'IA

Le point commun du machine learning et du deep learning est d'exploiter des données pour l'entraînement de modèles probabilistes. Avec les algorithmes/logiciels et le matériel, les données sont la troisième composante clé de la plupart des IA du jour.

Il faut donc se poser la question du lien entre la qualité des solutions d'IA et celle des données qui les alimentent.

On distingue généralement trois types de données pour entraîner un système de machine learning et de deep learning : les données d'entraînement, les données de test et les données de production.



Dans le machine et le deep learning supervisé, les données d'entraînement et de tests contiennent leur label, à savoir, l'information qui doit être générée par le système à entraîner. C'est un jeu de test doté d'une bonne répartition de l'espace du possible de l'application.

On le découpe arbitrairement en deux sous-ensembles, l'un pour l'entraînement et l'autre pour les tests de qualification du réseau de neurones entraîné qui déterminent un taux d'erreur de reconnaissance. En général, la part de la base taggée dédiée à l'entraînement est plus grande que celle qui est dédiée aux tests et dans un rapport 3/4 et 1/4.

Les données d'entraînement et de tests sont indispensables pour la grande majorité des systèmes d'IA à base de machine learning, que ce soit pour de l'apprentissage supervisé ou non supervisé. L'apprentissage par renforcement utilise une plus faible quantité de données mais s'appuie en général sur des modèles déjà entraînés au préalable.

On pourrait y ajouter les données de renforcement qui servent aux apprentissages par renforcement. On peut considérer qu'il s'agit de nouveaux jeux de données d'entraînement qui permettent d'ajuster celui d'un réseau de neurones déjà entraîné.

Données d'entraînement

Ce sont les jeux de données qui vont servir à entraîner un modèle de machine learning ou de deep learning pour en ajuster les paramètres. Dans le cas de la reconnaissance d'images, il s'agira d'une base d'images avec leurs tags corrrespondants qui décrivent leur contenu.

Plus la base est grande, meilleur sera l'entraînement du système, mais plus il sera long. Si vous n'avez pas de données déjà taggées pour entraîner un modèle de machine learning ou de deep learning, vous n'irez pas bien loin !

Les bases d'entraînement d'images ont une taille qui dépend de la diversité des objets à détecter. Dans l'imagerie médicale, des bases d'entraînement de pathologies spécialisées peuvent se contenter de quelques centaines de milliers d'images pour détecter quelques dizaines ou centaines de pathologies. A l'autre extrémité de la complexité, la base d'entraînement d'images de Google Search s'appuie sur plusieurs centaines de millions d'images et permet la détection de plus de 20 000 objets différents.

L'entraînement d'un système de 50 000 images dure environ un quart d'heures dans des ressources en cloud mais cela dépend des ressources allouées de ce côté-là. Lorsque l'on passe à des centaines de millions d'images, il faudra des milliers de serveurs et jusqu'à plusieurs semaines pour l'entraînement !

Dans la pratique, les jeux d'entraînement de solutions de deep learning sont limités en taille par la puissance de calcul nécessaire. On peut aussi réaliser un entraînement incrémental au gré de l'ajout de données avec des techniques de transfert de réseau de neurones[194].

Il est évidemment nécessaire de disposer de données d'entraînement de qualité, ce qui nécessite souvent un gros travail de filtrage, de nettoyage et dédoublonnage préalable à l'ingestion des données, une tâche existant déjà dans le cadre d'applications de big data.

Données de test

Ce sont les données, également taggées, qui serviront à vérifier la qualité de l'entraînement d'un système. Ces données doivent avoir une distribution statistique voisine des données d'entraînement, au sens où elles doivent être bien représentatives de la diversité des données que l'on trouve dans la base d'entraînement et que l'on aura dans les données de production.

Les données de tests sont un sous-ensemble d'un jeu de départ dont une partie sert à l'entraînement et une autre partie, plus limitée, sert aux tests. Elles seront injectées dans le système entraîné et on en comparera les tags résultants avec les tags de la base. Cela permettra d'identifier le taux d'erreur du système. On passera à l'étape suivante lorsque le taux d'erreur sera considéré comme acceptable pour la mise en production de la solution.

Le niveau de taux d'erreur acceptable dépend de l'application. Son maximum généralement accepté est le taux d'erreur de la reconnaissance humaine. Mais comme on est généralement plus exigeant avec les machines, le taux véritablement accepté est très inférieur à celui de l'Homme.

[194] Voir par exemple Incremental Learning in Deep Convolutional Neural Networks Using Partial Network Sharing, 2017 (12 pages).

Données de production

Il s'agit des données non taggées qui alimenteront le système lors de son utilisation en production pour faire une prévisions des tags manquants.

Alors que les données d'entraînement sont normalement anonymisées pour l'entraînement du système, les données de production peuvent être nominatives ainsi que les prévisions associées générées par la solution.

La nouvelle réglementation RGPD de l'Union Européenne exige que les entreprises conservent les données personnelles des utilisateurs ainsi que les données générées. Cela concerne donc à priori les données générées par les systèmes à base d'IA. Une donnée personnelle générée artificiellement reste une donnée personnelle ! Et son origine artificielle doit être connue et traçable en cas d'audit.

Données de renforcement

J'utilise cette expression pour décrire les données qui servent à l'apprentissage par renforcement. Dans un chatbot, cela sera par exemple les données de réactivité des utilisateurs aux réponses des chatbots permettant d'identifier celles qui sont les plus appropriées.

En quelque sorte, ce sont des résultats d'A/B testing réalisés sur les comportements d'agents à base d'IA. Tout ce qui pourra être capté sur la réaction du monde réel aux agissements d'un agent à base d'IA permettra potentiellement d'en ajuster le comportement par réentraînement.

L'apprentissage par renforcement est finalement une sorte d'apprentissage supervisé incrémental car on l'utilise pour faire évoluer par petites touches impressionnistes des systèmes déjà entraînés.

Origine des données

Les données alimentant les systèmes d'IA proviennent de l'intérieur et/ou de l'extérieur de l'entreprise.

Elles sont issues de toutes sortes de capteurs divers : des objets connectés, du plus simple (thermomètre connecté) aux plus sophistiqués (machine outil, smartphone, ordinateur personnel). Comme pour les applications de big data habituelles, les sources de données doivent être fiables et les données bien extraites et préparées avant d'être injectées dans les systèmes à base de machine comme de deep learning.

Les solutions les plus avancées exploitent conjointement des données ouvertes externes et les croisent aux données que l'entreprise est seule à maîtriser. C'est un bon moyen de créer des solutions différenciées.

Les données ouvertes sont issues de l'open data gouvernementale, des réseaux sociaux et de différents sites spécialisés dans la fourniture de données, soit ouvertes, soit payantes, comme des bases de prospects d'entreprises ou de particuliers, selon les pays et législations en vigueur. Côté images, il y a par exemple la base **ImageNet** déjà vue, la base **LSUN** qui contient des points de vues d'extérieur, **CIFAR-10** et **CIFAR-100**, des bases d'images diverses basse résolution et **Celeba**, qui contient 200 000 photos de visages de célébrités. Dans le langage, il y a notamment la base lexicale **WordNet** (anglais) avec ses 117 000 expressions et **MNIST** (écriture manuscrite).

La labellisation est un des défis majeurs de l'adoption rapide du machine learning par les entreprises[195]. Des sociétés spécialisées emploient des gens pour labelliser les données,

Les données d'entraînement des systèmes d'IA doivent être bien labellisées, soit automatiquement, soit manuellement. De nombreuses bases de référence d'images taggées l'ont été via de la main

[195] Voir This CEO is paying 600,000 strangers to help him build human-powered AI that's 'whole orders of magnitude better than Google', de Matt Weiberger, octobre 2018, et Could data costs kill your AI startup?, de Ivy Nguyen, novembre 2018.

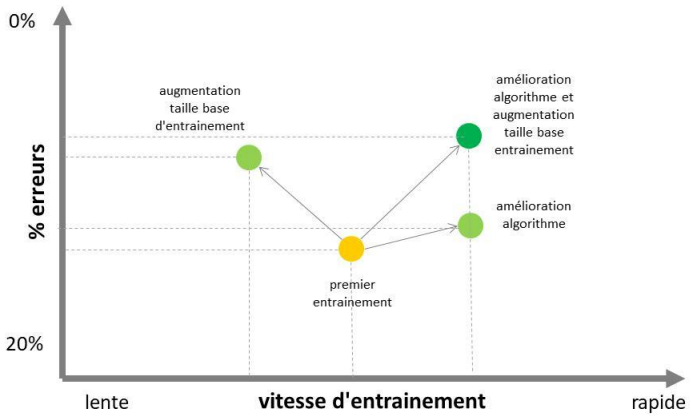
d'œuvre recrutée en ligne via des services du type d'**Amazon Mechanical Turk**[196] ou de **FigureEight** (2007, USA, $58M, aussi appelé CrowdFlower). Ces petites mains sont en quelque sorte les ouvriers de l'IA ! D'autres comme **thresher.io** (2015, USA), **OpenRefine** (open source) et **trifacta** (2012, USA, $124,3M) proposent des produits de collecte semi-automatique de labels afin de soulager les entreprises de la charge de labellisation. Cette labellisation peut cependant être de mauvaise qualité car pour des applications industrielles, une expertise métier est souvent nécessaire a une labellisation correcte. Le bon sens ne suffit souvent pas. Pour y pallier, **Neuromation** (2017, USA, $25M) et **Deep Vision Data** (USA) proposent de générer automatiquement des données d'entraînement numériques. C'est notamment utile pour créer des vues d'objets 3D en 2D sous plusieurs angles de vue pour entraîner des réseaux convolutionnels de vision[197]. Les cas d'usage sont cependant restreints[198]. Certains experts estiment que cette friction constitue une motivation supplémentaire à l'extension des usages du deep learning et de l'apprentissage par renforcement[199].

Là encore, les entreprises et les startups devront prendre en compte le règlement européen RGPD dans la collecte et le traitement des données personnelles. Leur portabilité d'un service à l'autre sera l'une des obligations les plus complexes à gérer. Le droit à l'oubli également[200] !

Certaines études portant sur un seul type de réseau de neurones montrent qu'une IA avec plus de données est plus efficace qu'une IA avec un meilleur algorithme.

La performance des algorithmes joue cependant un rôle clé dans la qualité des résultats dans le deep learning, et surtout dans leur performance, notamment la rapidité de la phase d'entraînement des modèles. Pour ce qui est de la reconnaissance des images, il faut distinguer le temps d'entraînement et le pourcentage de bonnes reconnaissances. Les progrès des algorithmes visent à améliorer l'une comme l'autre. La taille des jeux de données est en effet critique pour bien entraîner un modèle.

Si l'algorithme utilisé n'est meilleur que dans la vitesse d'entraînement, ce qui est souvent le cas dans des variantes de réseaux de neurones convolutionnels, alors, la performance de la reconnaissance ne changera pas lors de l'exécution du modèle entraîné. Par contre, avec plus de données d'entraînement, celui-ci sera plus long. Comme illustré dans mon petit schéma *ci-dessus*, il faut à la fois de meilleurs jeux de données et de meilleurs algorithmes pour que l'entraînement soit aussi rapide que possible.



[196] Voir Inside Amazon's clickworker platform: How half a million people are being paid pennies to train AI, de Hope Reese et Nick Heath, 2016.

[197] Voir Some startups use fake data to train AI, de Tom Simonite, avril 2018 et Deep learning with synthetic data will democratize the tech industry, de Evan Nisselson, mai 2018.

[198] Voir Deep learning with synthetic data will democratize the tech industry, mai 2018.

[199] Voir 5 tips to overcome machine learning adoption barriers in the enterprise, de Alison DeNisco Rayome, novembre 2017.

[200] Lorsqu'un réseau de neurones aura été entraîné avec des données personnelles de millions d'utilisateurs, la suppression des données personnelles d'une base de données ne signifiera pas automatiquement qu'elles ont disparu du réseau de neurones entraîné avec. Mais les données utilisées dans l'entraînement sont normalement anonymisées puisqu'elles servent à déteminer des caractéristiques des utilisateurs à partir de paramètres divers (localisation, comportement, usages). Les données ont beau être anonymisées, elles figurent sous la forme d'influence probabiliste du réseau de neurones entraîné. Influence qui est normalement négligeable à l'échelle d'un seul utilisateur. A l'envers, un réseau de neurones bien entraîné peut deviner des caractéristiques cachées d'un client via son approche probabiliste. Ces informations déduites doivent donc probablement être aussi bien protégées que les informations d'origine sur l'utilisateur.

C'est notamment utile pour réduire la consommation énergétique de l'IA. Bref, pour faire de l'IA verte[201] ! Mais il est bon de tenir compte des délais du projet : l'optimisation d'un algorithme pourra prendre beaucoup plus de temps que son alimentation avec un jeu de données plus grand.

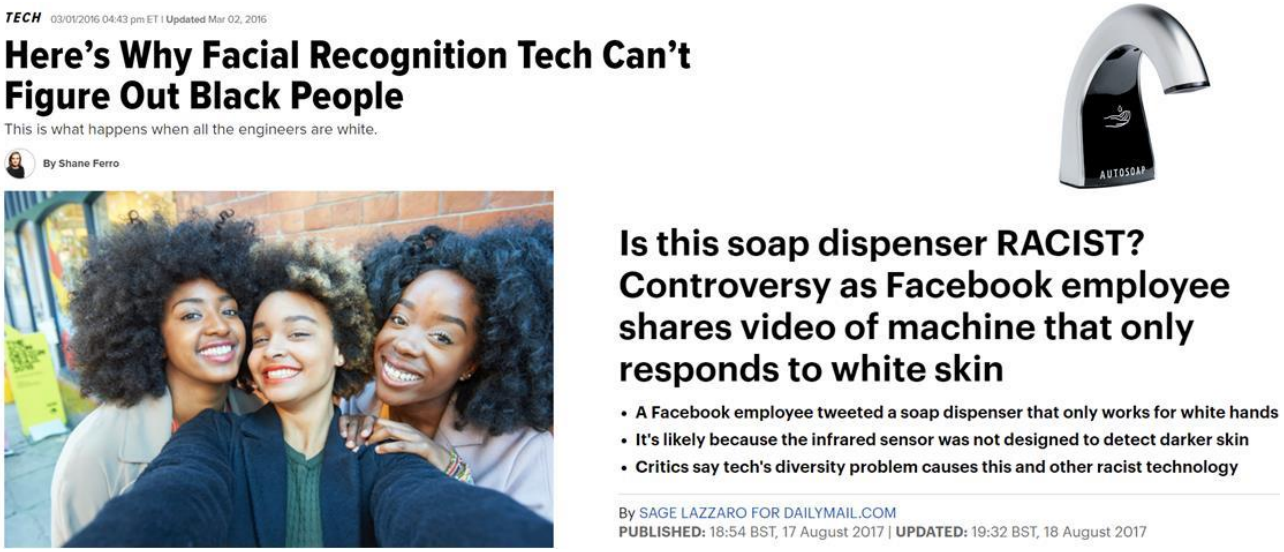
Où ces données sont-elles stockées ? Elles peuvent l'être sur les serveurs de l'entreprise ou dans le cloud et si possible dans un cloud bien privé de l'entreprise. Contrairement à une idée répandue, les services de cloud issus des GAFAMI n'exploitent pas les données des entreprises qui y sont stockées. Seules celles qui proviennent des services grand public (moteurs de recherche, réseaux sociaux, email personnels) peuvent l'être.

Par contre, les données qui circulent sur Internet peuvent être interceptées par certains services de renseignement qui ont installé des sondes sur les points d'accès des grandes liaisons intercontinentales. La DGSE le fait pour les fibres qui arrivent en France et la NSA pour celles qui arrivent aux USA, en général à des fins de renseignement sur le terrorisme mais cela peut déborder sur d'autres besoins !

Biais des données

Le biais des algorithmes est souvent évoqué car il peut affecter les résultats des traitements de machine learning et de deep learning. Mais le biais le plus fort est celui des données qui les alimentent[202].

Deux anecdotes l'illustrent parfaitement : chez Facebook, les femmes ingénieures de couleur se sont rendu compte en 2016 que les détecteurs de main dans les distributeurs de savon dans les WC ne fonctionnaient pas avec elles. Pour ces mêmes personnes, certains systèmes de reconnaissance de visages ne fonctionnent pas mieux. Pourquoi donc ?



Dans le premier cas, cela peut-être lié au capteur utilisé. Dans le second, c'est une histoire de données d'entraînement qui ont alimenté le système de reconnaissance de visage. Le point commun : les créateurs de ces systèmes n'avaient pas de personnes de couleur dans leurs équipes techniques. D'où un biais dans le matériel, dans les logiciels et les données.

---

[201] C'est un des objectifs du chercheur **Stéphane Canu** de l'INSA Rouen qui planche sur l'optimisation de gros modèles de vision artificiel et de traitement du langage. D'où le projet de recherché collaborative "Deep in France" lancé par différents laboratoires et financé par l'ANR.

[202] Voir l'ouvrage de référence Weapons of Math Destruction de Cathy O'Neil et cette vidéo d'une heure où elle résume son propos ainsi que son interview Les algorithmes exacerbent les inégalités, novembre 2018.

Dans la reconnaissance de visages, le biais vient de ce que les données comprenaient une proportion trop faible de visages de couleur. Et il semble que cela persiste dans les systèmes d'aujourd'hui selon une étude du MIT[203].

Une IA doit donc être alimentée par des jeux de données d'entraînement qui sont les plus représentatives des usages à couvrir, et notamment en termes de diversité d'utilisateurs. Cela demande de l'empathie, cela exige pour les créateurs de ces solutions de sortir de leur cadre de vie habituel[204].

En termes statistiques, cela veut dire que les données doivent avoir un fort écart type et une distribution similaire à celle du marché visé[205]. Les données d'entraînement d'IA qui portent sur le fonctionnement de machines doivent répondre aux mêmes exigences. Bref, pour faire du machine learning, il faut conserver de solides notions de statistiques et probabilités et ne pas se faire embobiner par des promesses douteuses[206].

En juillet 2018, l'assocation **ACLU** (American Civil Liberties Union) fit un test de croisement entre les 535 membres du Congrès US et 25 000 visages d'une base de criminels en open date avec le SDK Rekognition d'Amazon[207]. Le système reconnu 28 criminels de la base dans les 535 élus ! Amazon s'est défendu en mettant en avant la configuration utilisée qui tolérait 20% d'erreurs alors qu'ils recommandent à la police d'utiliser un seuil de 5%. On est ici dans un cas de biais de configuration d'algorithme plus que de données.

Ainsi, si on entraîne une IA à reconnaître le bruit de moteurs en panne, il faut disposer d'une base d'entraînement de bruits de moteurs représentative des divers types de pannes qui peuvent affecter les-dits moteurs. Sinon, certaines pannes ne seront pas détectées en amont de leur apparition.

Un autre biais peut affecter le machine learning lorsque celui-ci est affecté à la réalisation de prévisions: le biais des « **données-rétroviseur** ». En effet, les données du passé ne correspondent pas à l'avenir. Vous en êtes les témoins au quotidien lorsque vous planifiez un voyage dans un pays ou cherchez un produit donné, puis faite votre achat. Dans les jours et mois qui suivent, vous pourrez alors être bombardé de reciblage publicitaire sur ces mêmes pays et produits alors que vous n'en avez plus besoin. Tous les achats ne relèvent pas de « repeat business » !

---

[203] Voir Study finds gender and skin-type bias in commercial artificial-intelligence systems, de Joy Buolamwini dans MIT News, février 2018 (video).

[204] Voir Forget Killer Robots—Bias Is the Real AI Danger de John Giannandrea (Google), octobre 2017,

[205] En juin 2018, IBM annonçait la diffusion à venir d'une base contenant un million de visages en open data, avec une bonne représentation de la diversité. Elle serait cinq fois plus grande que la plus grande base ouverte actuelle. La base sera complétée d'une base encore mieux labellisée de 36 000 visages représentant de manière identique l'ensemble des ethnies, ages et genres. Voir https://www.ibm.com/blogs/research/2018/06/ai-facial-analytics/.

[206] L'illustration de potirons illustre de manière imagée ce besoin de diversité dans les données. Elle est extraite d'une présentation de Nikos Paragios lors de la conférence FranceIsAI d'octobre 2018 à Station F, à Paris.

[207] Voir Amazon's Facial Recognition Matched Congress Members to Criminals, juillet 2018.

On en a eu une autre belle démonstration avec cette prévision des résultats de la Coupe du Monde FIFA 2018[208] réalisée en juin 2018 par Goldman Sachs. Elle anticipait une victoire du Brésil contre l'Allemagne en finale et une défaite de la France contre le Brésil en demi-finale. Tout en indiquant que par essence, le football était un sport difficile à prédire. Alors, pourquoi donc faire des prévisions ? Qui se sont bien évidemment royalement plantées ! On pourrait même dire qu'il était statistiquement hautement probable que ces prévisions soient à côté de la plaque tant la théorie du chaos règne dans une telle compétition[209].

Autre exemple connu, le cas de **Google Flu** qui tentait de prédire les épidémies de grippe en se basant sur les recherches portant sur le sujet. Cela fonctionnait bien en moyenne mais l'outil a loupé l'épidémie de 2013, notamment aux USA. Même si l'outil n'utilisait probablement pas de deep learning, il est probable que si cela avait été le cas, il lui aurait été difficile de prédire l'épidémie[210].

Une mésaventure du même type a été découverte chez **Amazon** dont le système d'assistance au recrutement avait tendance à défavoriser les femmes candidates car les données d'entraînement comportaient très peu de femmes, comme c'est le cas dans de nombreux métiers techniques. La sous-représentation statistique des femmes dans le modèle avait donc tendance à diminuer leurs chances d'être retenues comme candidates. Cela montre que si l'on souhaite que les IA ne perpétuent pas des situations existantes insatisfaisantes, il faut modifier les représentations statistiques des données d'entraînement pour aller vers un équilibre meilleur ou plus équitable selon les critères du moment[211]. Là encore, il s'agit d'éviter l'effet rétroviseur des données du passé.

Une entreprise qui dépendrait trop des données du passé pour prédire le futur est l'analogue d'un conducteur de véhicule dont le rétroviseur aurait la taille du pare brise. Le platane serait difficile à éviter dans ces conditions ! C'est bien évidemment métaphorique mais requiert de prendre des précautions. Ce biais du rétroviseur est celui qui conduit les IA à être sexistes ou racistes, parce que les contenus utilisés pour l'entraîner reflètent le monde réel… qui l'est [212]! Finalement, le défaut des IAs est de trop refléter le monde existant et pas assez le monde souhaitable ! Et ce monde souhaitable ne peut être obtenu qu'en injectant un biais explicite dans le choix des données, pas une sorte de biais subi sans avoir réfléchi.

C'est une question autant politique que technique. Il en va de même pour des applications éminentes de l'IA que sont les agents vocaux, assez souvent féminins (Alexa, Cortana, …), les robots sexuels ou les armes robots. Ici, les données n'y sont pour rien. Il s'agit de choix sociologiques, marketing et politiques de leurs créateurs[213].

Dans l'ouvrage **Prediction Machines**, Ajay Agrawal, Joshua Gans et Avi Goldfarb[214] indiquent ainsi que les entreprises devront faire preuve de discernement pour allouer des tâches de prévisions aux IAs ou aux humains.

Les biais d'incomplétude interviennent couramment dans le traitement du langage lorsque des expressions sont interprétées mot à mot au lieu de l'être comme idiomatismes.

---

[208] Voir Goldman Sachs used AI to simulate 1 million possible World Cup outcomes — and arrived at a clear winner, juin 2018.

[209] Voir Coupe du Monde : le big data s'est encore spectaculairement raté dans ses prévisions, de Sylvain Rolland dans La Tribune, juillet 2018 qui fait état d'autres prévisions qui étaient tout autant à côté de la plaque que celle de Goldman Sachs.

[210] Voir Deep Learning: Diminishing Returns? de Bernard Murphy, juillet 2018.

[211] Voir Pourquoi il faut défendre Amazon et son algorithme de recrutement d'Aurélie Jean, dans le Point, octobre 2018.

[212] Voir Les intelligences artificielles sont-elles sexistes ? Des spécialistes nous répondent de Mathilde Saliou, février 2018.

[213] Voir AI has a gender problem. Here's what to do about it, de Samir Saran et Madhulika Srikumar, avril 2018.

[214] Cités dans The Economics Of Artificial Intelligence - How Cheaper Predictions Will Change The World dans Forbes en juillet 2018.

Une manière de visualiser cela est d'imaginer ce qu'un réseau de neurones génératif créérait comme image pour les expressions familières « mettre les pieds dans le plat » et « parler la langue de bois », *ci-contre*. Les métaphores ne sont pas encore le fort de l'IA !

L'autre biais humain classique est de confondre corrélation et causalité. Les exemples abondent dans ce sens comme cette vague corrélation entre la consommation de chocolat dans un pays et son nombre de prix Nobel par habitant.

Dans un tel cas, il existe au moins une demi-dizaine d'autres critères qui influent ces deux paramètres. Le machine learning entraîné uniquement avec des données de consommation de chocolat et de prix Nobel n'y verra que du feu.

Par contre, une analyse multiparamètres sera peut-être pertinente. Ici, c'est le choix de la structure des données analysées qui influera le résultat.

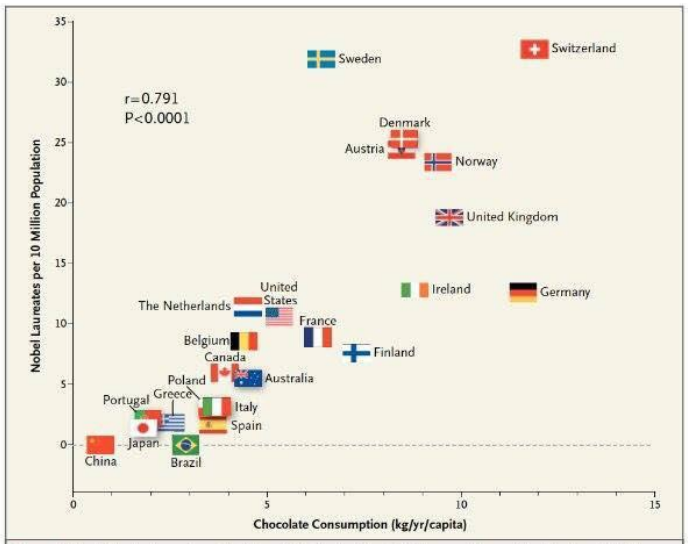


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Capteurs et objets connectés

Les capteurs et objets connectés jouent un rôle clé dans de nombreuses applications d'intelligence artificielle. Les micros et caméras alimentent les systèmes de reconnaissance de la parole et de vision artificielle. Les smartphones et les outils d'accès à Internet en général créent des tomberaux de données sur les comportements des utilisateurs. La smart city et les véhicules autonomes sont aussi alimentés par moult capteurs en tout genre.

L'un des moyens de se rapprocher et même de dépasser l'homme est de multiplier les capteurs sensoriels. La principale différence entre l'homme et la machine réside dans la portée de ces capteurs. Pour l'homme, la portée est immédiate et ne concerne que ses alentours. Pour les machines, elle peut-être distante et globale. On voit autour de soi, on sent la température, on peut toucher, etc. Les machines peuvent capter des données environnementales à très grande échelle. C'est l'avantage des réseaux d'objets connectés à grande échelle, comme dans les "smart cities". Et les volumes de données générés par les objets connectés sont de plus en plus importants, créant à la fois un défi technologique et une opportunité pour leur exploitation.

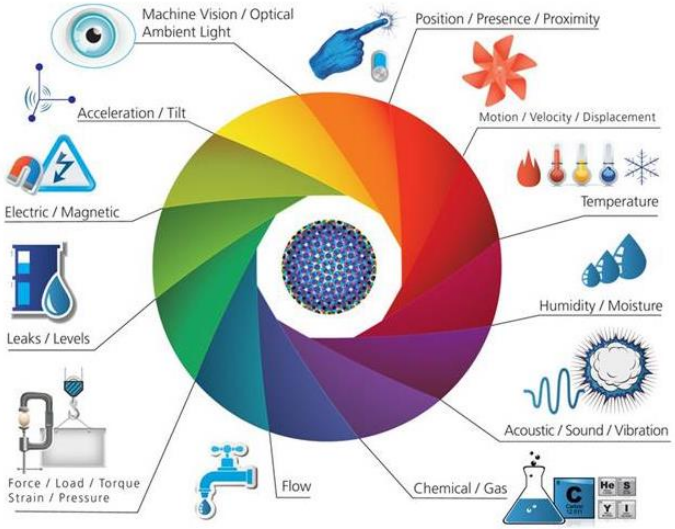
Le cerveau a une caractéristique méconnue : il ne comprend pas de cellules sensorielles. Cela explique pourquoi on peut faire de la chirurgie à cerveau ouvert sur quelqu'un d'éveillé. La douleur n'est perceptible qu'à la périphérie du cerveau. D'ailleurs, lorsque l'on a une migraine, c'est en général lié à une douleur périphérique au cerveau, qui ne provient pas de l'intérieur. L'ordinateur est dans le même cas : il n'a pas de capteurs sensoriels en propre. Il ne ressent rien s'il n'est pas connecté à l'extérieur. Une IA sans capteurs ni données ne sert à rien.

Cette différence peut même se faire sentir à une échelle limitée comme dans le cas des véhicules à conduite assistée ou automatique qui reposent sur une myriade de capteurs : ultrasons, infrarouges, vidéo et laser / LIDAR, le tout fonctionnant à 360°. Ces capteurs fournissent aux ordinateurs de bord une information exploitable qui va largement au-delà de ce que le conducteur peut percevoir, surtout dans la mesure où les données de plusieurs capteurs sont combinées (« sensor fusion »).

C'est l'une des raisons pour lesquelles les véhicules autonomes sont à terme très prometteurs et plus sécurisés. Ces techniques sont déjà meilleures que les sens humains, surtout en termes de temps de réponse, de vision à 360° et de capacité d'anticipation des mouvements sur la chaussée (piétons, vélos, autres véhicules). A contrario, la finesse de la vue humaine n'est pas encore égalée par la vision artificielle de fait de ses contraintes actuelles.

En effet, les réseaux de neurones convolutionnels utilisent des images sources à basse résolution pour tenir compte des contraintes matérielles actuelles. Ils sont rares à fonctionner en 3D avec une vision stéréoscopique[215].

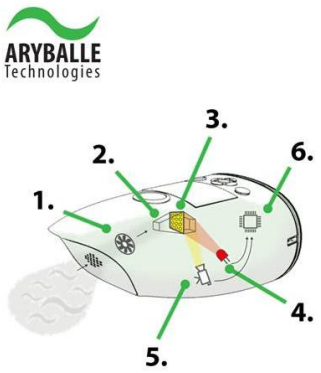
Le marché des capteurs a connu un fort développement depuis la fin des années 2000 grâce à l'émergence du marché des smartphones, alimenté par l'iPhone et les smartphones Android. Il s'en vend actuellement environ 1,5 milliards d'unités par an et ils sont renouvelés à peu près tous les deux ans par les consommateurs.



N'importe quel smartphone comprend au minimum une douzaine de capteurs : deux à quatre caméras, un à deux micros, un accéléromètre, un gyroscope, un GPS, un capteur de lumière, un capteur de proximité et des capteurs radio Bluetooth / Wifi / 2G / 3G / 4G.

Cela a eu comme conséquence d'accélérer la miniaturisation et la baisse du prix de tous ces capteurs. Les innovations dans le secteur des capteurs se poursuivent à un bon rythme et permettent de créer des perceptions extra-sensorielles par rapport aux capacités humaines. Chacun de ces capteurs va générer des données exploitables par des systèmes de machine learning et deep learning pour comparer le signal acquis avec des bases de données de signaux déjà associés à des matières déjà détectées.

Nous en avons deux exemples avec les spectrographes infrarouges comme ceux de l'israélien **Scio**, intégrés dans une balance de **Terraillon** ou un smartphone de Changhong lancé au CES 2017 (mais qui ne semble pas avoir percé ni été renouvelé en 2018), avec le détecteur de gaz Neose du français **Aryballe** ou encore avec le détecteur de pollution aérienne d'un autre français, **Plume Labs**.



[215] Voir un exemple dans Usings CNNs to Estimate Depth from Stereo Imagery de Tyler S. Jordan er Skanda Shridhar, 2014 (6 pages) ainsi qu'avec 3D Facial Expression Reconstruction using Cascaded Regression 2018 (8 pages).

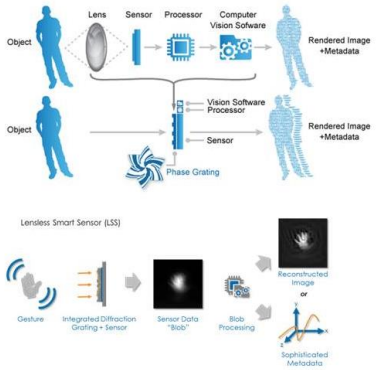
Les plateformes de gestion de maisons connectées tirent aussi parti de nombre de capteurs d'ambiance pour optimiser le confort. Ils jouent sur l'intégration de données d'origine disparate : la température extérieure et intérieure, l'humidité, la luminosité ainsi que les déplacements des utilisateurs, captés avec leur smartphone. Cela permet par exemple d'anticiper la température du logement en anticipation du retour au domicile de ses occupants.

Cette orchestration passe de plus en plus souvent par de l'apprentissage profond pour identifier les comportements des utilisateurs et adapter les réponses du système.

L'innovation dans les capteurs photo et vidéos est également incessante, ne serait-ce que par la miniaturisation de ceux qui équipent les smartphones et sont maintenant dotés de vision en 3D. L'américain **Rambus** planche de son côté sur un capteur photo qui n'a pas besoin d'optique ! Les capteurs de vibrations et les microphones ont des applications industrielles insoupçonnées et révélées par l'IA : la détection d'anomalies.

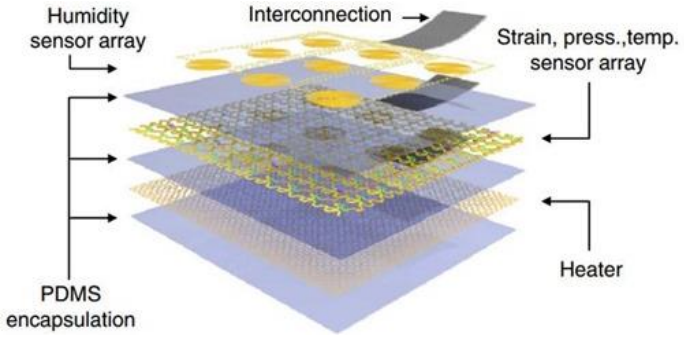
Rambus

capteur photo sans optique



Ainsi, des capteurs placés dans des véhicules ou des machines industrielles génèrent un signal qui est analysé par des systèmes de deep learning capables d'identifier et caractériser les anomalies. Ainsi, la société **Cartesiam** (France, 2,5M€) installée à Angers depuis 2016 a créé une slutions logicielle intégrée pour capteurs de vibration intégrant un réseau de neurones servant à détecter les vibrations anormales[216].

Les capteurs de proximité intégrables à des machines comme les robots progressent même dans leur biomimétisme. Des prototypes de peau artificielle sensible existent déjà en laboratoire, comme en Corée du Sud (*ci-contre*, **source dans Nature**).



L'une des mécaniques humaines les plus difficiles à reproduire sont les muscles.

Ils restent une mécanique extraordinaire, économe en énergie, fluide dans le fonctionnement, que les moteurs des robots ont bien du mal à imiter.

Les **capteurs biométriques** sont de plus en plus courants : dans les bracelets type FitBit, dans les montres connectées avec leurs capteurs infrarouges détectant le pouls et l'oxègénation du sang et dans les capteurs d'électroencéphalogrammes (EEG). Ces derniers permettent à l'homme de contrôler un membre artificiel robotisé, une application pouvant restaurer des fonctions mécaniques de personnes handicapées, voire démultiplier la force de personnes valides, dans les exosquelettes dédiés aux applications militaires ou dans le BTP. L'homme peut ainsi piloter la machine car la périphérie du cortex cérébral contient les zones où nous commandons nos actions musculaires.

Les caméras dans le visible et l'infrarouge couplées à d'éventuels autres capteurs permettent de détecter l'état psychologique de personnes à distance, comme leur niveau d'intérêt dans une conférence ! C'est un cas d'usage de la société française **datakalab** qui propose cela dans les conférences et même pour les utilisateurs d'Internet, en captant les émotions visuellement et via un bracelet connecté.

[216] Eolane et Cartesiam marient capteur et intelligence artificielle pour la maintenance prédictive des équipements industriels, avril 2018.

Matériel de l'IA

Le matériel est la troisième roue du carrosse de l'IA après les algorithmes, les logiciels et les données. Son rôle était méconnu jusqu'en 2016 lorsque l'on a commencé à évoquer le rôle des GPU puis des processeurs neuromorphiques. Depuis, l'industrie des semiconducteurs s'est mobilisée dans la course à l'IA. Je l'avais notamment constaté en préparant le Rapport du CES 2018.

Nous allons ici creuser le rôle et le fonctionnement de ces divers processeurs et de ce qui les accompagne côté mémoire et stockage. Leurs évolutions respectives contribuent aussi à améliorer la qualité et la performance des solutions d'intelligence artificielle.

L'un des outils clés de l'IA sont les serveurs d'entraînement du deep learning. Si celui-ci donne de très bons résultats, comme dans la reconnaissance d'images, il est très consommateur de ressources dans sa phase d'entraînement. Il faut facilement 1000 à 100 000 fois plus de puissance machine pour entraîner un modèle de reconnaissance d'images que pour l'exécuter ensuite.

Cela explique pourquoi, par exemple, les GPU et autres TPU (Tensor Processing Units) ont une capacité de calcul d'environ 100 Tflops/s tandis que les briques neuronales des derniers Kirin 980 de Huawei et de l'A11 Bionic se contentent de 1 à 4 Tflops/s. Et encore, l'entraînement des plus gros réseaux de neurones réclame la mise en batterie de centaines voir de milliers de serveurs utilisant ces GPU et TPU.

Processeurs

La loi de Moore est la pierre angulaire de nombreuses prédictions technologiques, notamment pour ce qui concerne celles de l'intelligence artificielle. Présentée comme immuable et quasi-éternelle, cette loi empirique indique que la densité des transistors dans les processeurs double tous les 18 à 24 mois selon les versions. Elle est aussi déclinée à foison pour décrire et prédire divers progrès techniques ou technico-économiques, y compris l'avènement de la fameuse singularité qui verrait l'IA dépasser l'intelligence de l'Homme.

La loi de Moore est aussi déclinée avec des variantes dans la vitesse des réseaux, la capacité de stockage, le cout d'une cellule solaire photovoltaïque ou celui du séquençage d'un génome humain. Une progression n'en entraîne pas forcément une autre. Le cout peut baisser mais pas la performance brute, comme pour les cellules solaires PV. On peut donc facilement jouer avec les chiffres.

Pour ce qui est de l'IA, on peut dire que la loi de Moore fait feu de tout bois. Elle se manifeste sous des formes des plus variées. Avec la course à la densité dans le CMOS qui commence à descendre en-dessous de 10 nm avec des roadmaps qui vont jusqu'à 3 nm, avec des technologies alternatives comme la photonique ou le quantique, puis avec des architectures différentes de processeurs optimisées pour l'entraînement et l'exécution de réseaux de neurones. C'en est au point où des processeurs sont adaptés à des types particuliers de réseaux de neurones, en gros avec les uns qui couvrent le traitement de l'image et les autres, celui du langage voir du bruit.

Enfin, l'IA sert aussi à conçevoir des chipsets avec des techniques élaborées d'organisation du layout des transistors et fonctions. C'est une évolution naturelle des outils de conception de chipsets[217].

[217] Voir notamment Startup JITX Uses AI to Automate Complex Circuit Board Design de Evan Ackerman, juillet 2918 et Using AI In Chip Manufacturing de Ed Sperling, août 2018.

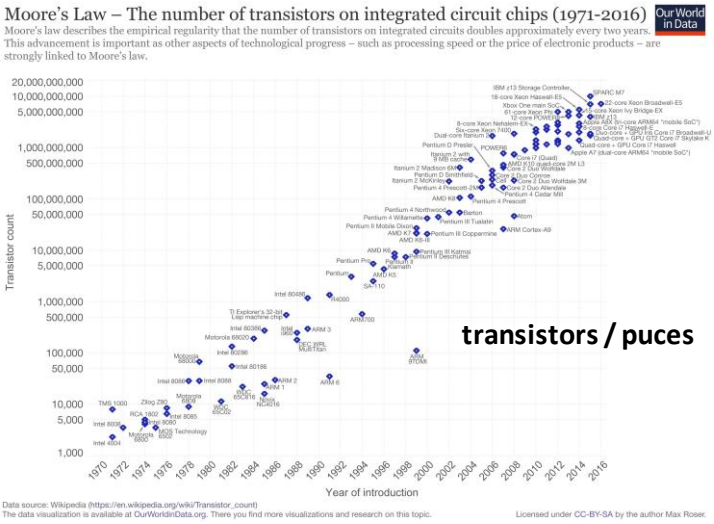
Course à la densité

Il fut un temps où tous les paramètres des chipsets CMOS évoluaient selon une loi exponentielle ou logarithmique selon les cas : leur densité, le nombre de transistors par chipsets, la vitesse d'horloge ou le prix par transistor.

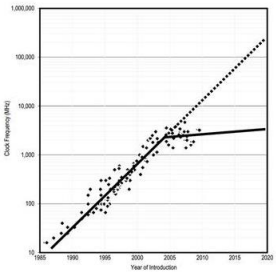
La loi de Moore est censée s'appliquer à des solutions commercialement disponibles, et si possible, en volume. Or ce n'est pas toujours le cas. Ainsi, l'évolution de la puissance des supercalculateurs est mise en avant comme un progrès technique validant la loi de Moore. Or, ces calculateurs sont créés avec des moyens financiers quasiment illimités et n'existent qu'en un seul exemplaire, souvent réalisé pour de la recherche militaro-industrielle ou de grands projets de recherche fondamentale (aérospatial, génomique, météo).

Dans la plupart des cas, ces technologies "de luxe" sont intégrées dans des produits grand public quelques années après. Ainsi, la puissance des super-calculateurs des années 1990 s'est retrouvée dans les consoles de jeu des années 2000. Au lieu de faire des calculs en éléments finis pour des prévisions météo, les consoles de jeux calculent des millions de polygones pour simuler des images en 3D temps réel. Mais cette puissance n'est pas homothétique dans toutes les dimensions. Si la puissance de calcul est similaire, les capacités de stockage ne sont pas forcément les mêmes.

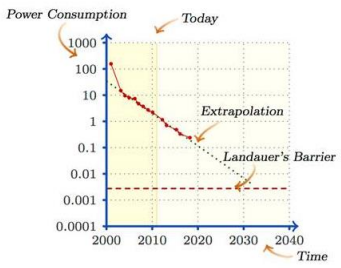
La densité des chipsets et leur nombre de transistors continue d'augmenter régulièrement (*ci-contre*) mais pas la fréquence d'horloge qui est au taquet aussi bien sur les laptops, desktops et serveurs. Pourquoi donc ? Pour ne pas faire fondre les chipsets et sur les mobiles, pour limiter la consommation d'énergie ! On plafonne ainsi à 2 GHz sur mobile et 4 GHz sur desktops et serveurs, sauf à utiliser des systèmes de refroidissement coûteux. D'où le choix des architectures multi-cœurs pour paralléliser les traitements qui sont exploitées des serveurs aux mobiles en passant par les ordinateurs personnels.



Par ailleurs, il existe une barrière assez méconnue, celle de Landauer, qui définit le niveau minimum d'énergie nécessaire pour gérer un transistor. Elle pourrait être atteinte d'ici 2035. Tout du moins, avec les architectures CMOS, mais pas avec celles de l'informatique quantique que nous verrons plus loin. Cette barrière est cependant contestée par certains physiciens.



barrière de la chaleur
limite la fréquence d'horloge des processeurs



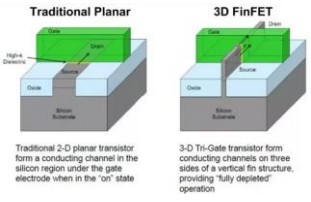
barrière de Landauer
limite basse de la consommation d'énergie d'un transistor

Les principales technologies d'intégration des 10 dernières années sont les transistors tri-gate « FinFET ». Avec eux, la densité est meilleure et il y a moins de fuites de courant et une baisse de consommation électrique.

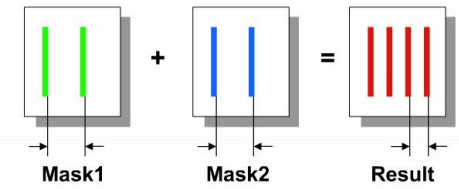
C'est encore mieux lorsque cette technologie est associée au FD-SOI issu du CEA-LETI, de Soitec et de STMicroelectronics, qui ajoute une couche d'isolant en oxyde de silicium (SiO2) sur les wafers de silicium et isole bien les transistors les uns des autres (voir le schéma *ci-dessous*).

Elle est notamment employée pour la fabrication de certains composants radio de l'iPhone ainsi que pour les chipsets neuromorphiques TrueNorth d'IBM. Le FD-SOI est aussi adopté sous licence par Samsung en 28 nm ainsi que par Global Foundries.

De plus, le multi-patterning[218] permet d'affiner le dessin des transistors mais il coute cher car il ajoute de nombreuses étapes à la fabrication des chipsets et peut augmenter le taux de rebus. La gravure en EUV (Extreme Ultra Violet) permet d'avoir des dessins plus nets et donc d'éviter ce multi-patterning.



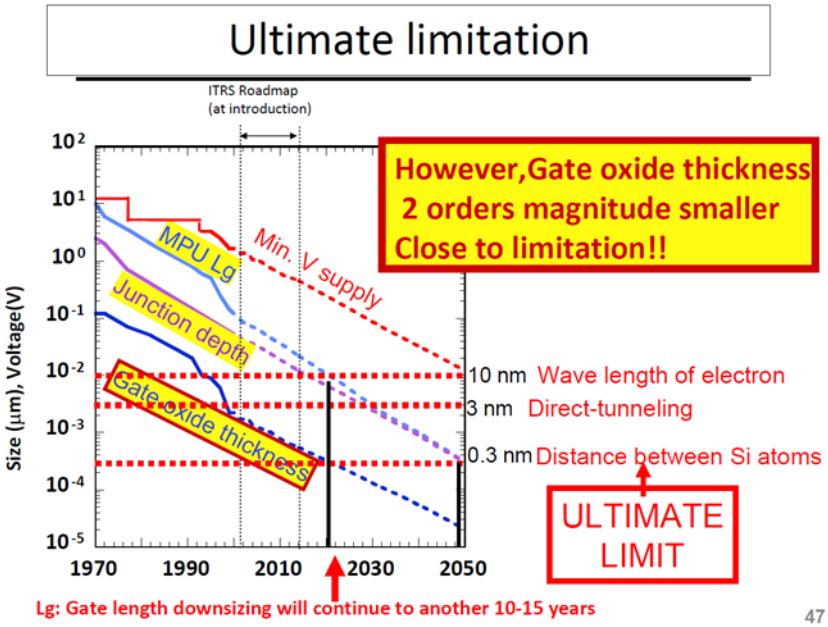
trigate transistors
less leaks, power saving
more complex steps



double and triple patterning
avoids small lines overlaps
but more process steps

La course à la densité est surprenante. Il y a une dizaine d'années, les spécialistes considéraient qu'il n'y avait point de salut en-dessous de 20 nm[219].

Même si, dans Downsizing of transistors towards its Limit, Hiroshi Iwai du Tokyo Institute of Technology (79 slides) décrivait ces limites avec précision en 2009 (*ci-contre*), la limite ultime étant la taille d'un atome de silicium[220]. Or les premiers chipsets en technologie 7 nm commencent à être fabriqués chez Samsung[221] et TSMC pour les Qualcomm Snapdragon 855 et les Kirin 980 de Huawei/HiSilicon.



[218] J'explique le principe du multi-patterning dans A la découverte de la "fab" chez STMicroelectronics : 2, décembre 2014.

[219] En voici un exemple avec Why is CMOS scaling coming to an END ? de Nor Zaidi Haron et Said Hamdioui, 2006 (6 pages).

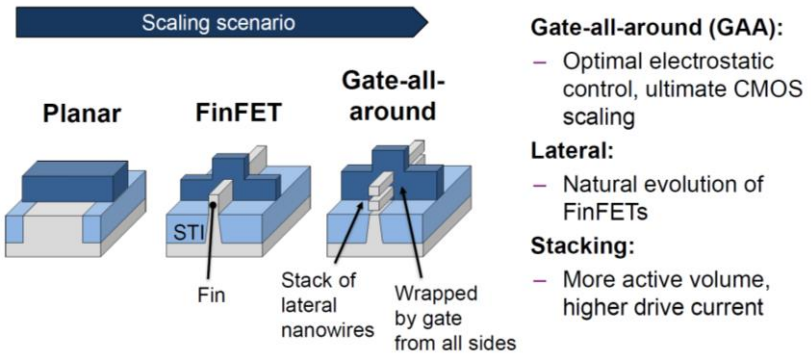
[220] L'excellent dossier **After Moore's Law**, paru dans The Economist en mars 2016, détaillait bien la question en expliquant pourquoi la loi de Moore des transistors CMOS pourrait s'arrêter en une douzaine d'année lorsque l'on descendra au niveau des 5 nm d'intégration.

[221] Voir Samsung Foundry Roadmap: EUV-Based 7LPP for 2018, 3 nm Incoming de Anton Shilov, mai 2018.

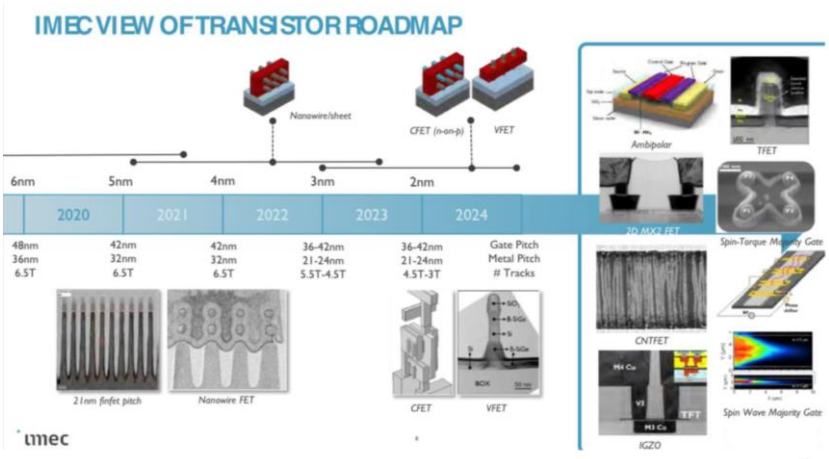
Et leurs roadmaps respectives descendent jusqu'au 3 nm d'ici une dizaine d'années ! A partir du 7 nm, la gravure en Extreme Ultra Violet deviendra indispensable du fait de sa meilleure résolution spatiale qu'elle procure.

Elle permettra au passage de limiter les étapes de patterning et de passer de plus de 100 à environ 80 masques[222]. Les machines d'**ASML** de gravure en EUV coutent environ $200M l'unité et il en faut environ 4 à 5 pour une unité de production classique [223] ! Cette montée en flèche des coûts explique les montants exhorbitants à investir pour des fabs <10 nm, de plus de $15B

L'une des techniques étudiées pour faire croître la densité des transistors est celle des « Gate-all-around » et des nano-fils qui permettent d'améliorer l'intégration verticale des transistors [224] . Le CEA-LETI est d'ailleurs en pointe sur le sujet avec sa technique CoolCube. La technologie ne semble pas encore industrialisée.



On retrouve cette technologie avec d'autres (TFET, CNTFET, etc) dans la roadmap Innovations Enabling Semiconductor Roadmap du Taïwanais Chee Wee Liu, 2018 (45 slides). Les roadmaps présentées descendent même à 2 nm d'ici 2024 grâce à l'emploi de ces transistors encore plus denses et intégrés verticalement. Pour mémoire, 1 nm = 10 atomes de silicium.

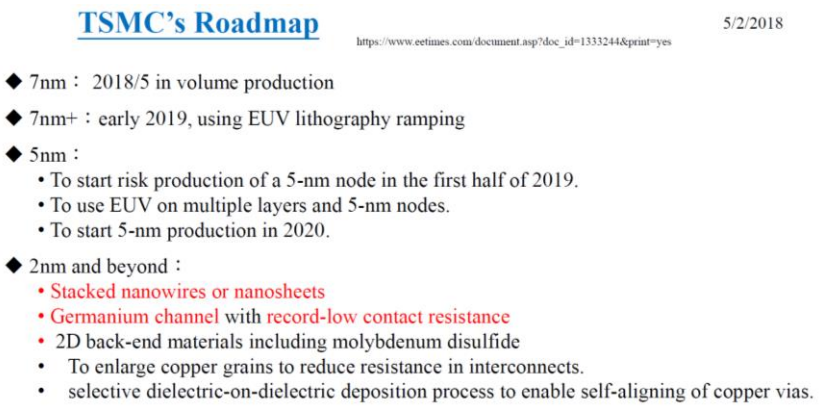


[222] Voir The Impact of EUV on the Semiconductor Supply Chain, 2018 (19 slides).

[223] Voir la description de la technologie et de la roadmap d'ASML dans Enabling Semiconductor Innovation and Growth - EUV lithography drives Moore's law well into the next decade, 2018 (37 slides).

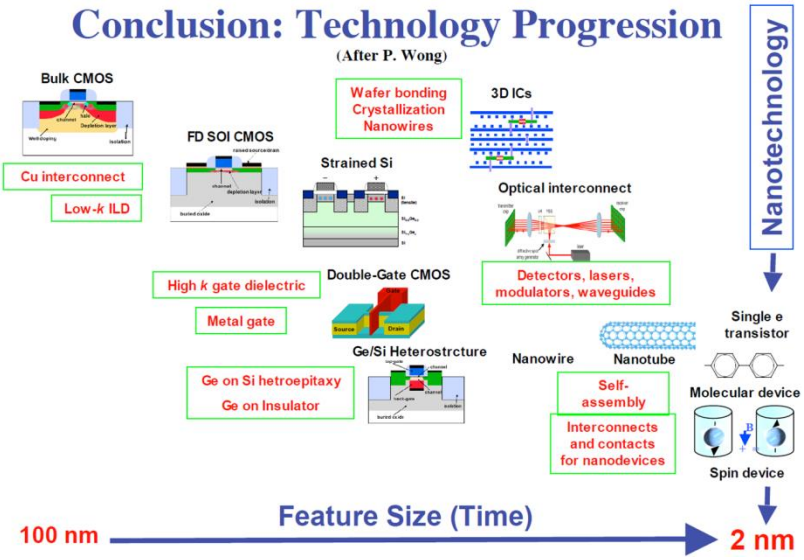
[224] Le schéma est issu de la présentation From Gate-all-around MOSFETS based on vertical stacked horizontal Nanowires de Hans Mertens, 2017 (78 slides).

Elle est aussi présente dans la roadmap du fondeur Taïwainais TSMC pour l'après 2 nm ! Ils voient bien loin ! Dans Performance and Design Considerations for Gate-All-around Stacked-NanoWires FETs, 2017 (47 slides) d'où est issu le slide *ci-contre*. TSMC fabrique notamment les chipsets d'Apple, Qualcomm et Nvidia



D'autres techniques sont envisagées à plus long terme comme des nanotubes de carbone comme chez **Nantero**[225] avec sa NRAM au standard DDR4 et proposée sous licence.

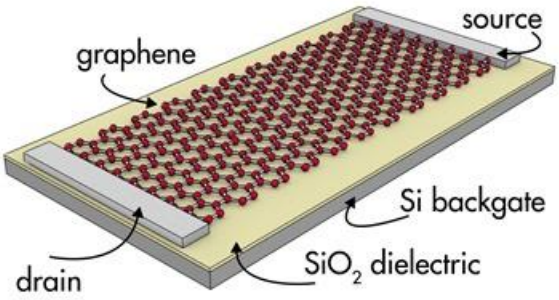
Le schéma *ci-contre* est issu de How far can we push Si CMOS What lies beyond de Krishna Saraswat, Stanford University (27 slides).



En-dessous de 2 nm, il faudra commencer à faire une croix sur la loi de Moore. Ou pas… !

Les architectures multi-cœurs atteignent de leur côté leurs limites car les systèmes d'exploitation et les applications sont difficiles à ventiler automatiquement sur un nombre élevé de cœurs, au-delà de 4.

Du côté de la vitesse d'horloge, des pistes sont explorées avec du graphène. IBM avait **annoncé** en 2011 avoir produit des transistors au graphène capables d'atteindre une fréquence de 155 GHz, et en gravure à 40 nm. Depuis, c'est le calme plat. Une performance en laboratoire d'aboutit pas toujours à de l'industrialisation ! Cela peut-être lié à une difficulté à fabriquer le composant avec un taux de défaut raisonnable.



Les laboratoires qui planchent sur le graphène depuis une dizaine d'année ont bien du mal à le mettre en œuvre en contournant ses écueils et à le fabriquer à un coût raisonnable. Il faudra encore patienter un peu de ce côté-là même si cela semble très prometteur et avec des débouchés dans tous les domaines et pas seulement dans l'IA.

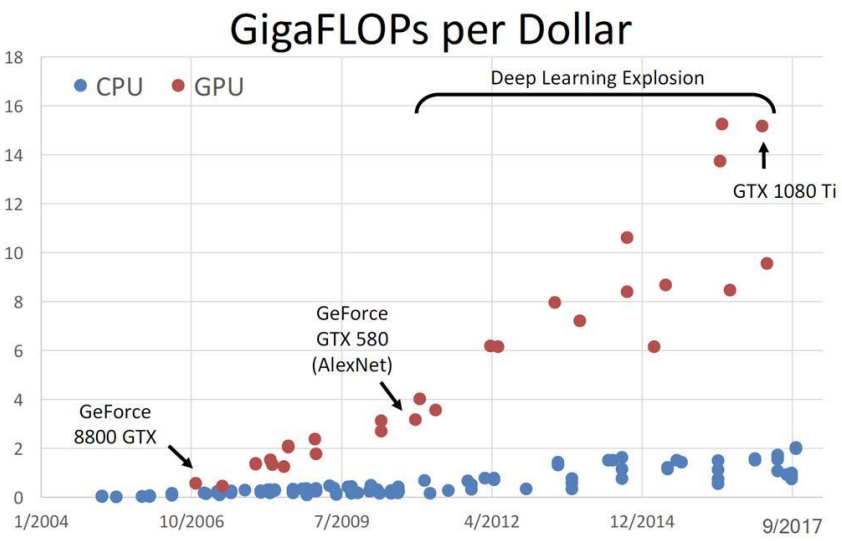
Alors, la loi de Moore est foutue ? Pas si vite ! Elle avance par hoquets et il reste encore beaucoup de mou sous la pédale pour faire avancer la puissance du matériel. Mais tout ce que nous venons de voir concerne principalement que la densité des processeurs. C'est loin d'être suffisant pour accélérer les logiciels de l'IA !

[225] Voir Carbon Nanotube DRAM, de Susan Rambo, août 2018.

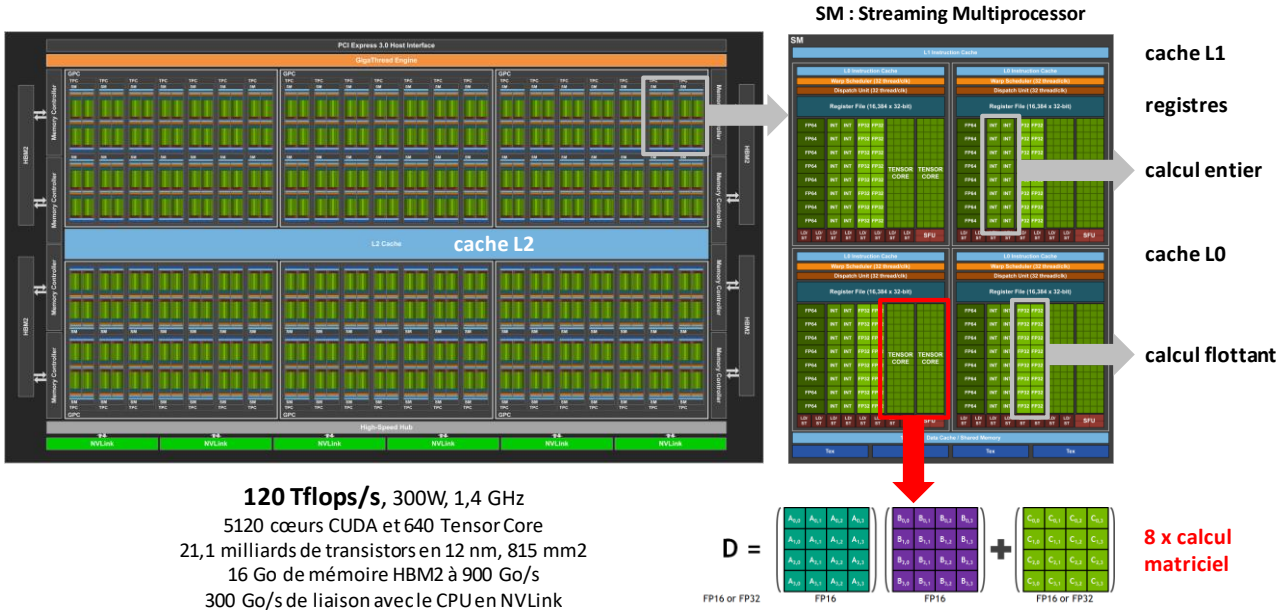
GPU

Les GPU constituent la solution matérielle la plus largement déployée pour accélérer les réseaux de neurones et le deep learning.

Ce sont eux qui ont rendu possible le deep learning, notamment pour le traitement de l'image et à partir de 2012. On le voit dans le schéma *ci-dessous* qui décrit la baisse drastique du coût au GigaFlops par dollar. La puissance brute a également augmenté d'autant sur le même laps de temps.



Le leader de ce marché est l'américain Nvidia qui fournit des chipsets et des cartes équipés de GPU dépassant les 5000 cœurs et complétés par des « Tensors Processing Units », à savoir des multiplicateurs de matrices. Ils sont couramment installés dans des data centers. Leur challenger AMD est à la peine, avec son API OpenCL qui bénéficie d'un support plus que médiocre par l'écosystème logiciel du deep learning[226].



Cette génération de GPU utilise des cœurs de génération Volta et totalise 21,1 milliards de transistors gravés en 12 nm. Ces GPU ont une puissance cumulée de 120 Téraflops/s[227] ! Jusqu'à présent, les GPU comprenaient une myriade de cœurs à même d'effectuer des opérations mathématiques simples (multiplications, divisions, additions, soustractions).

---

[226] Voir AMD Too Late To NVIDIA Deep Learning Party, de Boris Suvorov, décembre 2016. Et la dernière génération de cartes AMD, lancées en juin 2017, les Radeon Instinct M125, a une puissance théorique de 24,6 Tflops, à comparer aux dernières Nvidia V100 qui atteignent 120 Tflops.

[227] Voir pas mal de détails dans NVIDIA Volta Unveiled: GV100 GPU and Tesla V100 Accelerator Announced. A noter que Nvidia entretient une équipe de développeurs en France sous la responsabilité de Julien Demouth qui participe à la conception de ses GPU pour le deep learning.

Les logiciels utilisant l'interface CUDA répartissaient les traitements dans le GPU pour les paralléliser comme pour la génération des effets graphiques 2D et 3D. Pour le deep learning, les calculs étaient aussi répartis dans ces cœurs mais ce n'était pas optimal.

Avec les GV100, Nvidia a ajouté des « tensor cores », des multiplicateurs de matrices de 4x4 permettant de mieux paralléliser les traitements d'un réseau de neurones, surtout dans les réseaux convolutionnels.

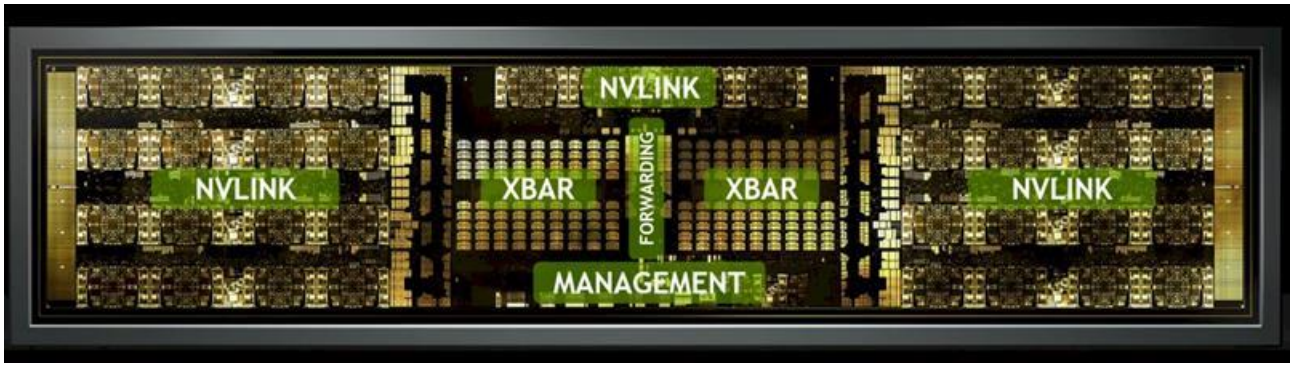
Ces GPU comprennent 80 « streaming multiprocessors », comprenant un total de 5120 cœurs CUDA traditionnels (avec 64 cœurs en flottant 32 bits, 32 cœurs flottant 64 bits et 64 cœurs entier par SM) et 640 « tensor cores » (8 par SM). Cette architecture présente l'avantage d'être assez flexible et générique et de s'adapter à de nombreux types de traitements. Elle est par ailleurs très bien supportée côté logiciels et frameworks.

L'architecture des Volta est finalement hybride : les ALU ne sont pas très utiles pour le deep learning car elles ne correspondent pas aux besoins d'optimisation aussi bien pour les réseaux convolutionnels liés au traitement de l'image qu'aux réseaux à mémoire dédiés au traitement du langage. Environ 30% de la surface du processeur contient ces multiplicateurs de matrices qui sont surtout utiles pour entraîner des réseaux convolutionnels.

Les V100 contiennent une mémoire au standard HBM2 de 16 Go avec un débit extraordinaire de 900 Go/s. Une version de 32 Go est sortie plus tard[228]. Cette quantité de mémoire est critique car elle conditionne le nombre de paramètres d'un réseau de neurones à entraîner. Or on atteint très vite cette quantité de mémoire avec un réseau convolutionnel de plusieurs couches. Cela explique entre autres choses pourquoi la résolution des images traitées dans ces réseaux de neurones est limitée à 256 ou 224 pixels de côté.

Comme l'explique **Tim Dettmers**, un GPU n'est utilisable pour des réseaux de neurones que si la mémoire est facilement partagée entre les cœurs de GPU. C'est ce que propose justement Nvidia avec son architecture **GPUDirect** RDMA et avec son bus NVLink qui atteint la vitesse de 300 Go/s avec ses derniers GPU GV100 Volta annoncés en mai 2017 et qui n'ont pas été renouvelés depuis lors.

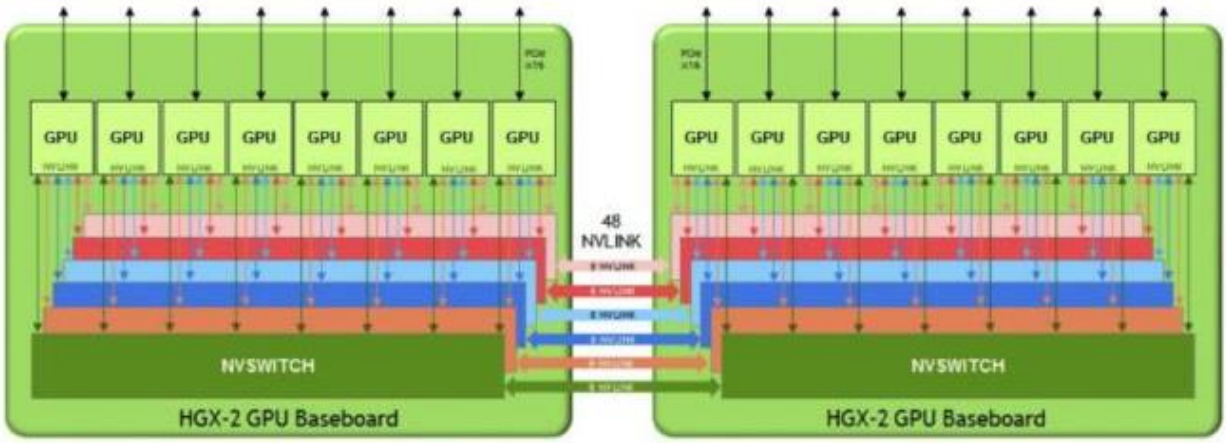
Les processeurs Volta équipent les serveurs DGX-1 par lots de 8 lancés en 2017 et les DGX-2 lancés en 2018 en comprenant 16, avec respectivement 1 et 2 petaflops de puissance théorique. Nvidia propose aussi les architectures de référence HGX-1 et HGX-2 destinée aux supercalculateurs hyperscale comme Microsoft Olympus et Facebook Big Basin. Nvidia propose aussi des versions station de travail des DGX, dotée de quatre cartes V100.



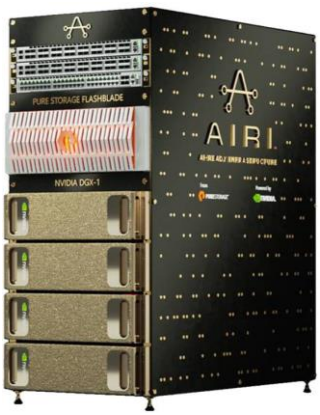
[228] Voir High-Performance Memory Challenges de Ann Steffora Mutschler, avril 2018, qui explique les défis de la gestion de la mémoire dans les GPU dédiés aux applications de l'IA.

Les DGX-2 sont des serveurs 10U (hauteur) équipés d'une nouvelle interconnexion NVLink 2.0 avec des chipsets dédiés, les NVSwitch (*ci-dessus*), avec 2 milliards de transistors, 18 ports NVLink 2.0 supportant 25 Go/s de communication bi-directionnelle par port donnant un total de 900 Go/s[229]. Le tout consomme 10KW et coute la bagatelle de $400K ! Autant dire que les développeurs y feront surtout appel via des opérateurs de cloud qui pourront investir dans ce genre d'engin.

En mai 2018, c'était le tour des HGX-2 d'être lancés. Voisine des DGX-2, elle comprend 16 GPU V100. Cette plateforme de référence sera utilisée par Lenovo, Supermicro, Wiwynn, QCT, Foxconn, Inventec, Quanta et Wistron.



Mais Nvidia a déjà un écosystème de partenaires dans le matériel. Ainsi **PNY** (1985, USA) propose AIRI (AI-Ready Infrastructure), un serveur d'IA prêt à l'emploi créé par Pure Storage et Nvidia. Il est dédié aux applications de machine learning et de deep learning. Ce système intègre quatre serveurs Nvidia DGX-1 fournissant 4 petaflops de puissance de calcul théorique, plus 17 To de stockage SSD, une double liaison fibre à 100 Gbits/s complété par les briques logicielles de Nvidia de deep learning et un « scaling toolkit » pour bien répartir les traitements sur les 32 GPU de l'ensemble.



En France, **Atos** a lancé en 2017 ses serveurs hybrides BullSequana S qui associent des CPU Intel, des GPU Nvidia et des extensions à base de FPGA, le tout permettant d'optimiser ses calculs dans l'IA.

En septembre 2018, Nvidia lançait une carte GPU au format PCIe de 75 Watts pour serveurs dédiée à l'exécution de réseaux de neurones entraînés et intégrant un GPU Tesla T4 exploitant l'architecture Turing apparue dans les GPU lancés au printemps. Ces cartes Tensor Hyperscale Inference Platform intègrent maintenant des serveurs de datacenters. Le GPU Nvidia Tesla T4 comprend 320 noyaux Turing Tensor Cores (multiplicateurs de matrices) et 2560 noyaux CUDA classiques.

[229] Voir Inside Nvidia's NVSwitch GPU Interconnect et Building biger, faster GPU clusters using NVSwitches tous deux publiés en avril 2018.

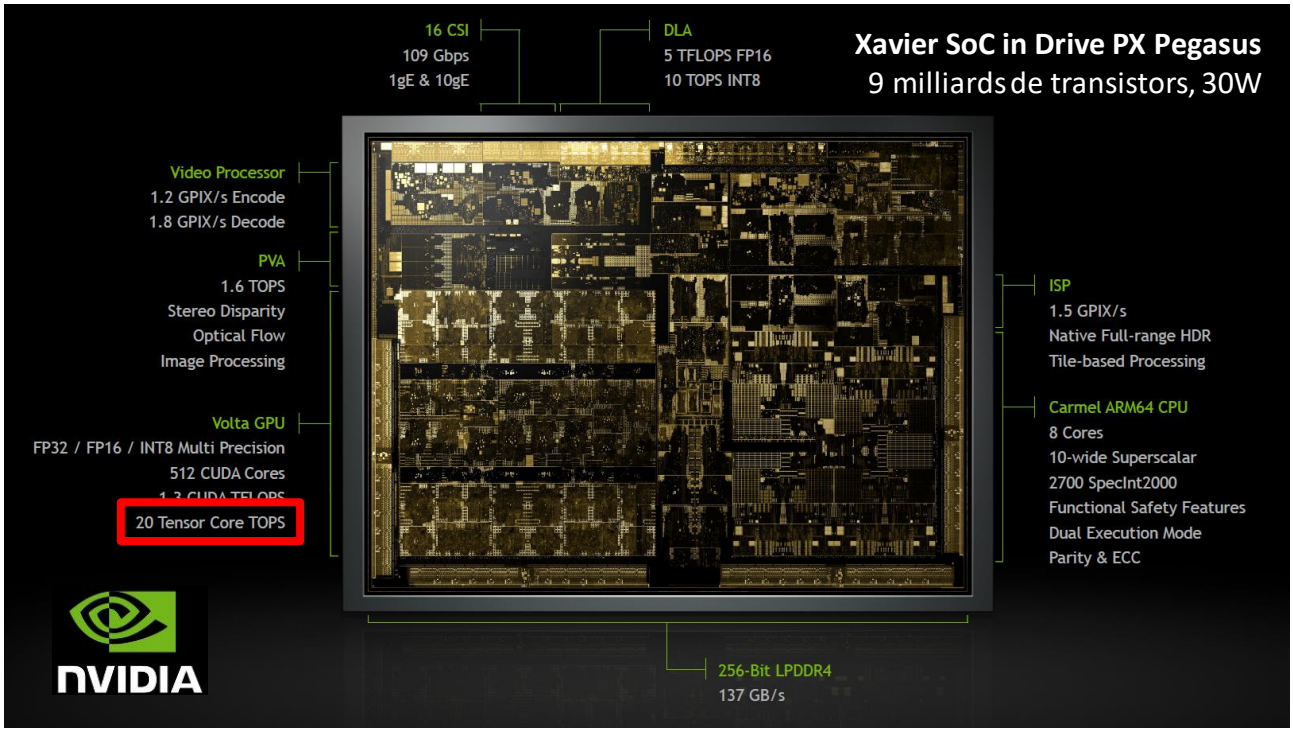
Les calculs sont réalisés avec des précisions FP16 et FP32 (16 et 32 bits en nombre flottant) ainsi que INT8 et INT16 (entier sur 8 et 16 bits). Il atteint 65 téraflops en FP16, 130 téraflops en INT8 et 260 téraflops en INT4[230]. Le GPU est supporté par la bibliothèque logicielle TensorRT 5 qui comprend un optimiseur d'inférence et un runtime supportant les frameworks TensorFlow, MXNet, Caffe2, Matlab et les frameworks supportés par ONNX comme ceux de Microsoft.

Nvidia lançait aussi en décembre 2017 une offre d'accès à ces serveurs en cloud dans le Nvidia GPU Cloud qui est proposée gratuitement aux chercheurs sous certaines conditions comme l'usage de différents GPU ou serveurs de la marque.

Nvidia propose aussi ses cartes **Jetson** pour l'informatique embarquée. L'architecture est complète, des objets connectés aux serveurs, permettant de répartir les traitements en fonction de leur nature ainsi que des réseaux de télécommunications utilisés.

Un système de surveillance peut ainsi disposer de son intelligence locale pour n'envoyer vers les serveurs que des alertes consommant peu de bande passante et via des réseaux télécoms de type LPWAN (Low Power Wide Area Network) comme celui de Sigfox.

En février 2018, **Nvidia** annonçait un partenariat avec **AnyVision** pour créer des caméras de surveillance intégrant de la reconnaissance faciale. Cela s'intègre dans l'initiative Metropolis de Nvidia qui est destinée aux Smart City qui associe également **Cisco**, l'intégrateur de systèmes de vidéosurveillance **Genetec**, le spécialiste du machine learning **Omni AI** et **MotionLoft** qui propose une solution logicielle de comptage de personnes et de véhicules dans l'espace public.



En octobre 2017, Nvidia lançait sa nouvelle carte supportant la conduite entièrement autonome de niveau 5, la Nvidia Drive PX **Pegasus**. Elle a une puissance de 320 teraflops/s. Elle exploite quatre processeurs embarqués dont deux de la série Xavier embarquant un GPU de la série Volta et totalisant 9 milliards de transistors (*ci-dessus*).

C'est un chipset différent du V100 car il est plus généraliste, comme un chipset de smartphone. Il ne comprend que 20 tenseurs contre 640 dans le V100 ce qui est normal car ce genre de chipset sert à l'inférence de réseaux de neurones et pas à leur entraînement.

[230] D'où la nuance lorsque l'on parle de téraflops !

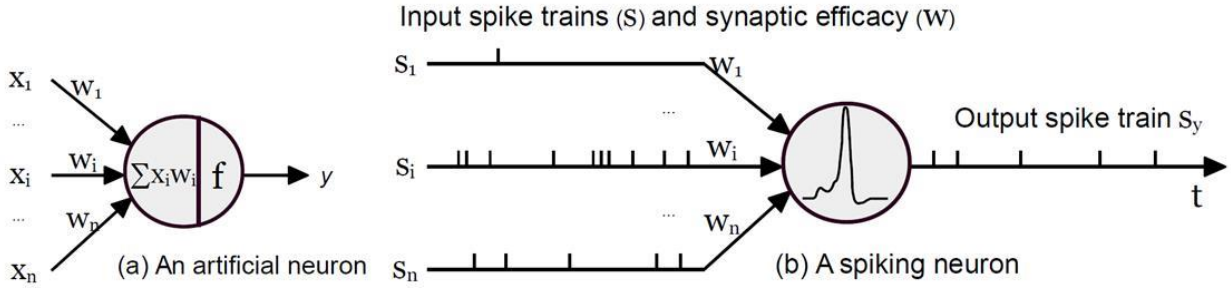
La carte s'interface avec 16 capteurs haut-débit donc des capteurs ultra-sons, caméras, radars et LiDARs. Elle est dotée de plusieurs connecteurs Etherner 10 Gbits/s. Sa bande passante mémoire excède 1 To/s. Cette carte permet l'exécution de modèles de deep learning entraînés sur des serveurs Nvidia DGX-1 ou DGX-2.

Processeurs neuromorphiques

Les processeurs neuromorphiques auraient été conceptualisés pour la première fois en 1990, par Carver Mead[231]. Ils sont spécialisés dans les traitements de machine learning et deep learning. Ils exécutent en parallèle les processus d'entraînement puis d'exécution d'applications de deep learning et leurs réseaux de neurones.

Il y a cependant de nombreuses variantes de tels processeurs neuromorphiques tant au niveau logique que physique[232]. Côté logique, certains d'entre eux sont plutôt adaptés à l'entraînement de réseaux convolutionnels avec des multiplicateurs de matrices et d'autres semblent conçus plutôt pour les réseaux à mémoire et le traitement du langage. Cela se traduit par des architectures physiques différentes qui sont plus ou moins proches de l'organisation de neurones physiques mises à plat dans leur layout[233].

Une autre piste est explorée qui consiste à imiter d'encore plus près le fonctionnement des neurones biologiques avec les systèmes à base de neurones à impulsion, ou *spiking neurons*[234]. L'un de ses intérêts est d'être très économe en énergie. Le neurone reçoit un train d'impulsions dans l'ensemble de ses synapses et génère en sortie un train d'impulsions résultat du calcul. Ces spiking neurones sont difficiles à programmer. Je n'ai d'ailleurs pas encore compris comment ils fonctionnaient par rapport aux réseaux de neurones récurrents ou temporels. C'est la voie choisie par IBM avec ses chipsets TrueNorth que nous verrons plus loin dans la partie dédiée aux memristors.



Les chipsets adaptés aux réseaux convolutionnels comprennent souvent des unités spécialisées dans la multiplication de matrices qui sont utilisées dans les premières couches de convolution. Les dernières couches « fully connected » s'appuient aussi sur des matrices reliant les neurones d'entrées avec les neurones des couches suivantes, via des grilles comprenant les poids des synapses. Mathématiquement, ce sont des unités de traitement qui multiplient des vecteurs à une dimension par des matrices pour générer des vecteurs.

---

[231] Voir Neuromorphic Electronic Systems, Carver Mead, 1990.

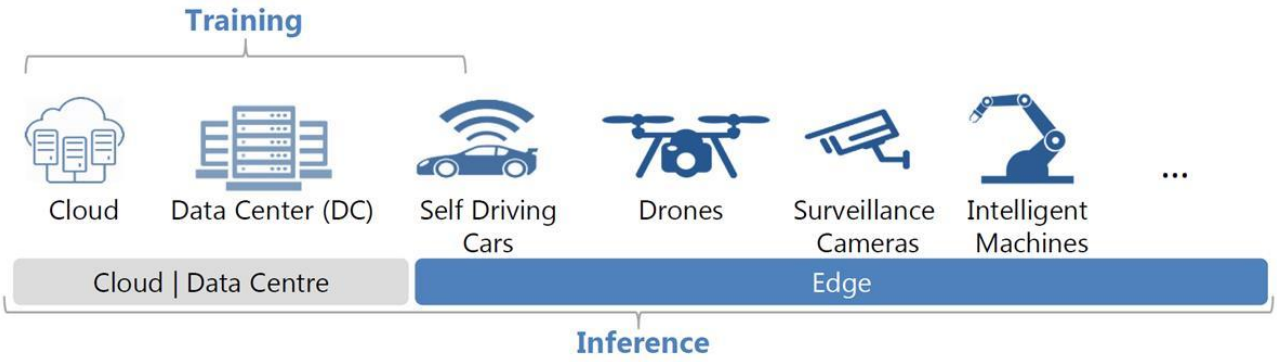
[232] Voir cet inventaire de différents types d'architectures de processeurs neuromorphiques : Neuromorphic Computing and Neural Network Hardware, 2017 (88 pages) qui s'appuie sur une bibliographie record de 2682 entrées. Vous pouvez aussi consulter l'excellente présentation Hardware Architectures for Deep Neural Networks, 2017 (290 slides).

[233] Voir quelques exemples de neurones plus ou moins proches des neurones biologiques : Brain synapses : Ultralow power artificial synapses using nanotextured magnetic Josephson junctions, janvier 2018 et Artificial Brain Synapses Replicated in a Chip, février 2018.

[234] Cette thèse Deep Spiking Neural Networks de Qian Liu, 2018 (212 pages) fait un point à jour de la technologie des spiking neurons. Elle porte sur leurs techniques d'entraînement, offline (sur d'autres architectures) et online (dans les spiking neurons). C'est de cette thèse que provient l'illustration de cette page avec la comparaison entre neurone artificielle classique et neurone à impulsion.

Ces chipsets sont aussi conçus pour que la mémoire qui stocke les poids des synapses et les *feature maps* des réseaux convolutionnels soit la plus proche des unités de traitement, afin d'en accélérer le fonctionnement, surtout pendant les phases d'entraînement[235], le cas le plus avancé étant celui des memristors.

L'autre variante se situe dans la position des chipsets dans la chaîne de valeur de l'IA. Certains chipsets très puissants sont adaptés à l'entraînement de réseaux de neurones sur des serveurs tandis que d'autres sont dédiés à l'exécution ou « inférences » de réseaux de neurones déjà entraînés dans des objets connectés ou smartphones. L'un des enjeux est pouvoir à terme faire de l'apprentissage par renforcement dans les chipsets embarqués.



Ces deux dimensions sont reflétées dans le schéma suivant de mon cru avec une bonne part des start-tups et sociétés citées dans ce qui va suivre. On y voit que l'offre est bien plus abondante du côté de l'embarqué que des serveurs. Cela vient notamment du fait que les composants côté embarqué sont souvent plus spécialisés alors que les chipsets pour serveurs sont plutôt génériques et coutent plus cher à conçevoir et à fabriquer.

[235] C'est une approche qui est aussi adoptée par la startup grenobloise **UpMem** (2015, France, $3,6M) qui conçoit des circuits de traitement intégrant mémoire et calcul (Processing-In-Memory ou PIM), mais dédiés au big data. Visiblement, l'architecture semble plus proche de celle des GPU que des processeurs neuromorphiques.

On peut aussi classifier ces processeurs suivant un autre axe lié à leur mode de réalisation physique :

- **FPGA** : ce sont des processeurs programmables qui sont créés pour de faibles volumes de production. On peut activer par logiciel les portes logiques du circuit pour créer des neurones sur mesure. Ils sont un peu l'équivalent de l'impression 3D pour les chipsets : intéressants pour les faibles volumes et le prototypage rapide mais moins pour les volumes importants. C'est la technologie retenue par **Microsoft** pour ses chipsets Brainwave qui exploite des FPGA d'origine Intel. On en trouve aussi chez diverses startups comme **Teradeep** (2014, USA) ou **Leapmind** (2012, Japon, $3,4M). Ces processeurs peuvent être 10 fois plus rapides que des GPU.
- **ASIC** : ce sont des chipsets fabriqués en volume dont le layout est défini une fois pour toute avant la fonderie. C'est la technique utilisée pour les processeurs Intel, pour les processeurs mobiles courants ou les GPU de Nvidia. Elle est adaptée aux gros volumes de production. Elle est aussi plus efficace côté puissance et économie d'énergie par rapport aux FPGA, pouvant aller jusqu'à un rapport de 1 pour 100 à 1000. C'est l'approche retenue par Google pour sa dernière génération de TPU[236].
- **Memristors** : ce sont des circuits de réseaux de neurones qui mémorisent de manière non volatile les poids des synapses des neurones au sein même des neurones. Dans les FPGA ou ASIC, ces informations sont stockées soit dans les circuits eux-mêmes, soit dans des RAM séparées, et de manière volatile. Le stockage de ce poids des neurones évite des accès mémoire externes et permet une accélération significative de la phase d'entraînement. Ces poids sont gérés sous forme de résistances variables non volatiles et donc sous forme analogique.

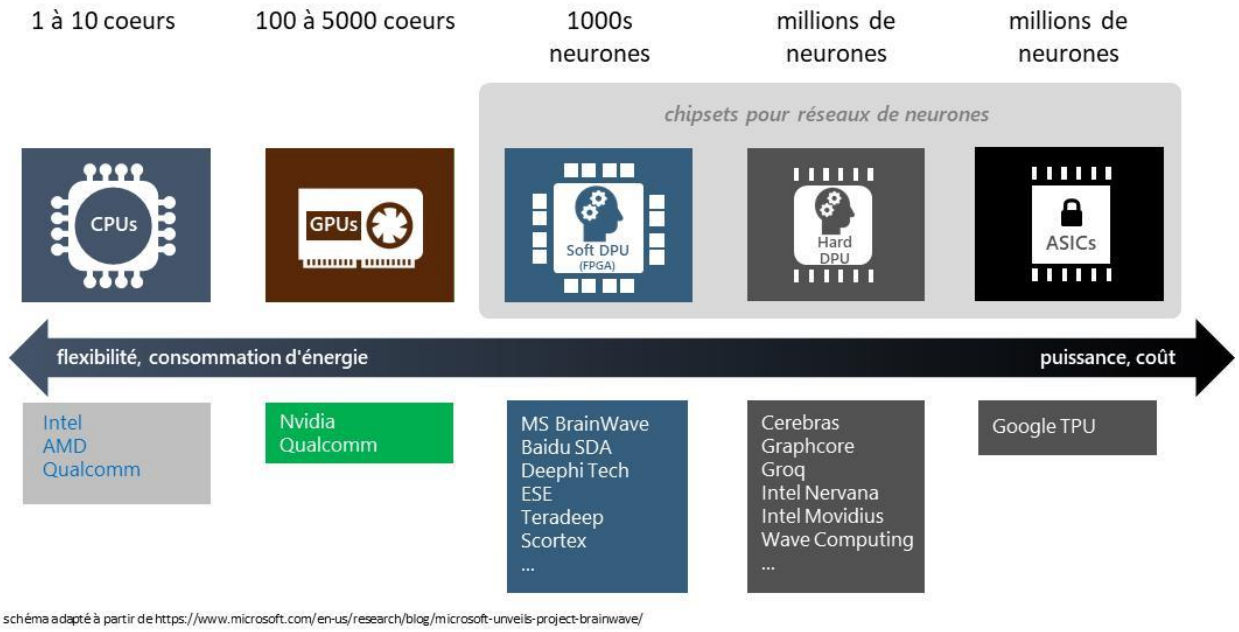


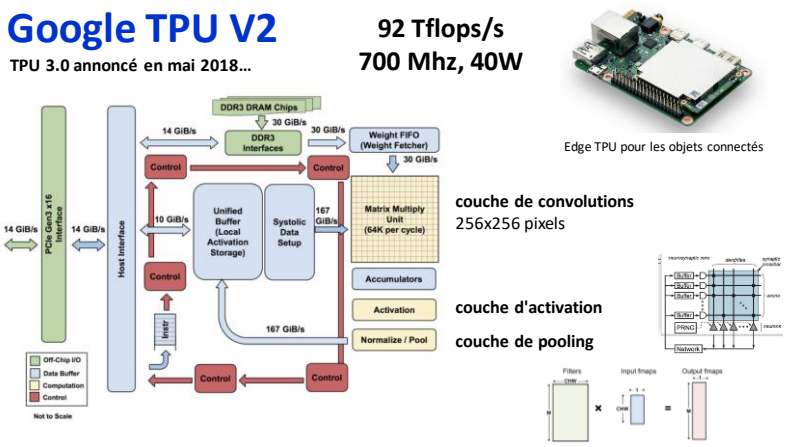
schéma adapté à partir de https://www.microsoft.com/en-us/research/blog/microsoft-unveils-project-brainwave/

Comme certains acteurs tels qu'Intel sont présents sur plusieurs créneaux à la fois, je vais ici décrire leur offre, fournisseur par fournisseur.

Google a créé ses TPU ou Tensor Processing Units adaptés notamment à l'exécution des applications développées avec TensorFlow. Ce sont eux qui ont permis la victoire d'AlphaGo au jeu de Go début 2016. Ils sont intégrés dans les datacenters de Google pour ses applications et services en cloud mais ne sont pas commercialisés séparément.

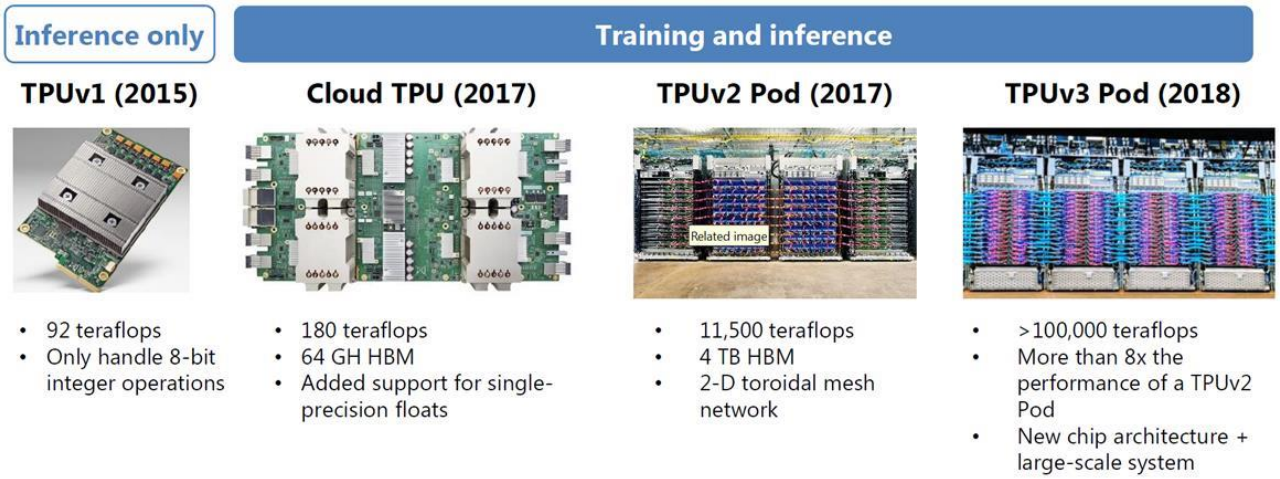
[236] L'approche de Google est décrite en détails ici : https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture/.

Ils en étaient à leur troisième génération à la mi-2018. Ils utilisent des ASIC performants et consommant peu d'énergie. Leur layout semble surtout adapté à l'exécution de réseaux convolutionnels avec une matrice de 256x256 pour gérer des convolutions ainsi que des couches de pooling (réduction de résolution) et d'activation (dernière couche de réseaux de neurones pour obtenir les tags des objets détectés)[237].



La troisième génération des TPU est 8 fois plus puissante que la seconde, sans que l'on dispose de plus de détails techniques autrement que leur capacité mémoire de 32 Go au standard HBM et le fait qu'ils doivent être refroidis par liquide. Les performances se chiffreraient en dizaines de pétaflops.

L'accès aux TPU est possible en cloud. D'après leur nomenclature, il semblerait que les TPU dénomment les racks de serveurs plutôt que les chipsets qu'ils contiennent[238]. Depuis 2017 (cf. schéma *ci-dessous*), les TPU servent aussi bien à l'entraînement de réseaux de neurones qu'aux inférences.



En juillet 2018, Google complétait le tableau avec ses cartes Edge TPU, des versions adaptées aux objets connectés, sans que les spécifications détaillées ne soient vraiment disponibles pour la puce TPU ssociée.

L'offre comprend une carte de développement équipée de chipsets à noyaux ARM et Vivante pour le GPU (Edge TPU Dev Board) et une carte USB équipée d'un TPU Google. Tous les deux sont adaptés à l'exécution d'inférences de réseaux de neurones développés avec TensorFlow Lite.

Intel a une offre pas très lisible mais très diversifiée en processeurs adaptés à l'entraînement et à l'exécution d'applications de deep learning dont une bonne part provient d'acquisitions.

---

[237] Voir des details sur l'architecture des TPU sur A Domain-Specific Architecture for Deep Neural Networks, septembre 2018.

[238] Voir cette description assez détaillée sur Tearing apart Google's TPU 3.0 AI coprocessor, mai 2018. Le schéma sur les TPU est issu de la presentation AI optimized chipsets de Vertex, 2018 (24 slides).

Nous avons en reprenant le schéma *ci-dessous* de gauche à droite :

- **Xeon** et **Xeon Phi** sont la gamme de processeurs généraliste dédiés aux serveurs. Intel fait des efforts pour optimiser les frameworks de deep learning (TensorFlow, Torch, Caffe) pour qu'ils s'exécutent plus rapidement sur des architectures Core et Xeon traditionnelles, alors qu'ils sont habituellement optimisés uniquement pour les GPUs type Nvidia. Cela aurait permis d'améliorer les performances d'un facteur x70 à x85 sur les processeurs Xeon[239] qui équipent les serveurs de data centers, rapprochant leurs performances des meilleurs GPU Nvidia de 2017. Bref, Intel aurait du mou sous la pédale dans ses processeurs serveurs !

- **Intel NNP-L** issu de l'acquisition de la startup Nervana en 2016 est une gamme de chipsets ASIC dédiée à l'entraînement et à l'exécution de réseaux de neurones sur serveurs. C'est vaguement l'équivalent des TPU de Google et des GPU V100 de Nvidia. L'offre Nervana était initialement intégrée dans Lake Crest[240] (2017) suivie dans la roadmap Intel par Knights Crest (circa 2020). Le coprocesseur Nervana embarque une mémoire au standard HBM2 de 32 Go permettant un transfert interne de données à la vitesse de 1 To/s, voisine des 900 Go/s du GV100 Volta de Nvidia. Intel n'est pas très bavard sur l'archirecture interne des tenseurs (multiplicateurs de matrices) de ses coprocesseurs !

- **FPGA Stratix 10** sont les FPGA en 14 nm d'Intel, issus de l'acquisition d'Altera en 2015[241]. Ces FPGA sont programmés par les clients d'Intel comme Microsoft le fait pour ses chipsets BrainWave. Ils présentent l'avantage d'intégrer des blocs de mémoire de 20 Kbits, utiles pour l'accélération de l'entraînement ou l'inférence de réseaux de neurones. Les Stratix 10 tournent en théorie jusqu'à 1 GHz, avec une capacité de 9,8 Gflops, et une puissance de 80 Gflops par Watt. Intel fournit son SDK OpenVino pour le développement d'applications de deep learning, notamment dans la vision.

- **Movidius VPU** correspond aux versions commerciales des chipsets Fathom de Intel Movidius annoncées mi 2017. Ces chipsets exploitant des processeurs vectoriels sont dédiés au traitement de l'image dans l'embarqué comme dans les caméras de surveillance. En août 2017, Intel annonçait une nouvelle génération de processeurs Myriad X, remplaçant les Myriad 2. Ces « Vision Processing Unit » destinées à l'embarqué permettent de traiter un trillion d'opérations de réseaux de neurones par secondes, soit 10 fois plus que pour les Myriad 2, grâce à 16 processeurs vectoriels au lieu de 12 et surtout, au passage côté fabrication à une architecture 16 nm vs 28 nm (chez TSMC, en ASIC). Au passage, le chipset peut aussi faire de l'encodage vidéo en 4K et ne consomme que 2W. Tout cela est issu de l'acquisition de la startup Fathom en 2016. Pour compléter tout cela, Intel faisait l'acquisition en août 2018 de la startup **Vertex.IA** (USA), un petite équipe de 7 personnes à l'origine de PLAID.ML[242], un framework open source de deep learning pour l'embarqué. Le chipset Movidius Myriad 2 équipe les nouvelles caméras infrarouges **FLIR** Firefly annoncées en octobre 2018. Ainsi, la détection d'intrusions est réalisée par interprétation directe des images dans la caméra. Il est aussi intégré dans les clés **USB Neural Compute Stick 2** qui peuvent être intégrées dans des serveurs, des PC ou des objets connectés.

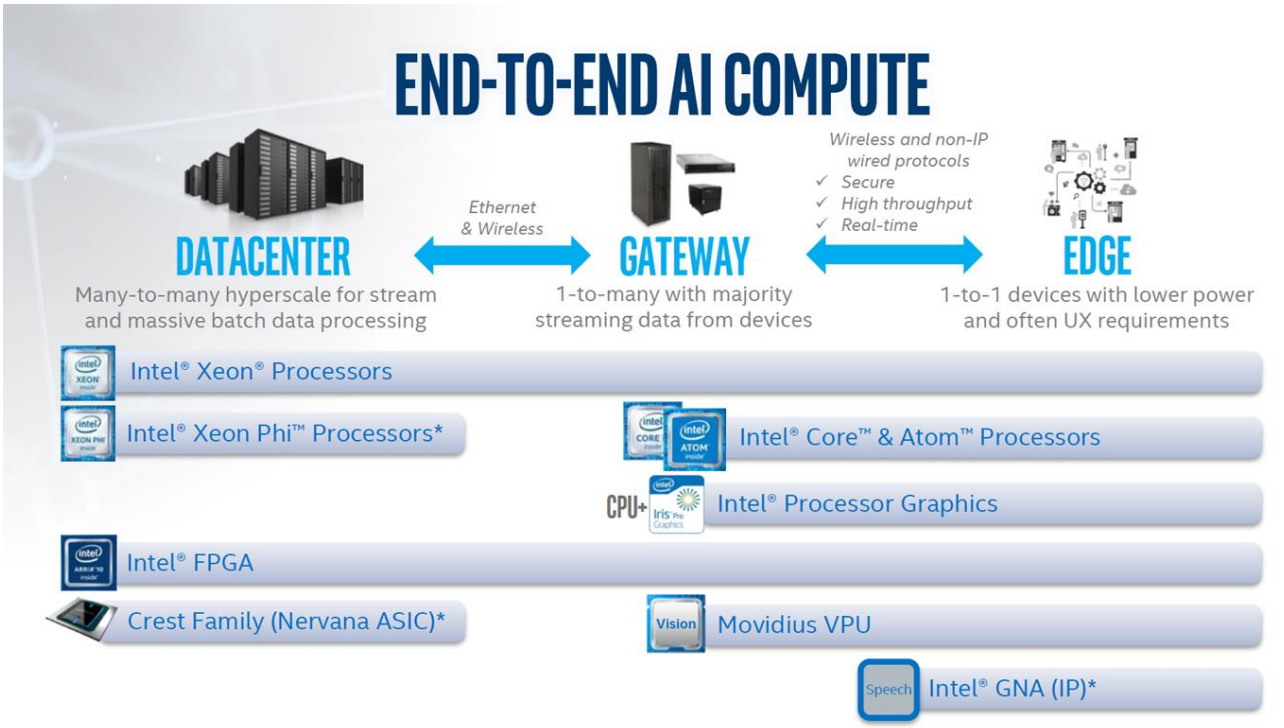
---

[239] Voir TensorFlow Optimizations on Modern Intel Architecture, août 2017 et New Optimizations Improve Deep Learning Frameworks For CPUs, octobre 2017. Ces optimisations s'appuient sur l'utilisation des instructions de traitements de vecteurs AVX2 des Xeon et AVX512 des Xeon Phi, ainsi que sur les versions 2017 des bibliothèques Intel Math Kernel Library (Intel MKL) et Intel Data Analytics Acceleration Library (Intel DAAL). Le jeu d'instruction AVX512 permet de réaliser des opérations matricielles voisines de celles des cœurs Tensor des TPU de Google et des GPU Nvidia GV100.

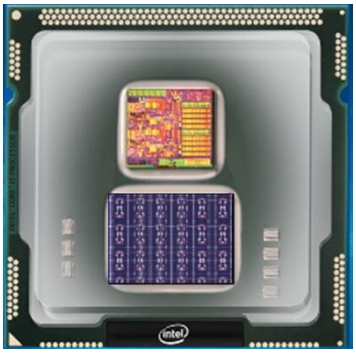
[240] Voir Intel Shipping Nervana Neural Network Processor First Silicon Before Year End, Anandtech, octobre 2017.

[241] Voir Machine Learning with Intel® FPGAs d'Adrian Macias, mai 2018 (32 slides).

[242] Il est sur GitHub ici : https://github.com/plaidml/plaidml. Le framework est particulièrement adapté aux ordinateurs tournant sous MacOS et Windows avec cartes graphiques AMD, Nvidia et les GPU intégrés dans les processeurs Intel. C'est en fait une surcouche qui supporte les applications développées avec le framework Keras ainsi qu'avec ONNX, Open Neural Network Exchange, un framework open source de bas niveau de réseaux de neurones, créé par Microsoft, Amazon et Facebook. Il supporte Caffe2, PyTorch, MXNet et le Microsoft CNTK. PLAID.ML est adapté aux réseaux de neurones convolutionnels (CNNs) et à mémoire (LSTM).



- **MobileEye** est une filiale d'Intel, issue d'une acquisition en 2017, qui a mis sur le marché depuis longtemps sont propre chipset embarqué pour la reconnaissance d'image pour les véhicules à conduite assistée et autonome exploitant une ou plusieurs caméras RGB classiques.
- **GNA** est un chipset dédié au traitement du langage servant de coprocesseur pour les processeurs des séries Atom et Core. C'est évidemment dédié aux systèmes embarqués.
- **Loihi** est un chipset annoncé en septembre 2017 et qui devait arriver à la mi-2018 sous forme d'un chipset de test pour la recherche. Devant être fabriqué en technologie 14 nm comme les Core i5/i7 de 2017/2018 et comprendre 130 000 neurones impulsionnels, comme dans les chipsets TrueNorth d'IBM avec des neurones reliés entre eux par 130 millions de synapses.

  En juin 2018, Intel montait les enchères en annonçant 100 milliards de synapses pour 2019, soit le niveau du cerveau d'un rat, en racontant des salades sur l'atteinte ultime de l'intelligence humaine[243]. Le marketing d'Intel évoque un processeur imitant le cerveau humain et doué de facultés d'apprentissage, en précisant qu'il supportera des réseaux de neurones récurrents, hiérarchiques et parcimonieux (sparse) et donc en particulier à tout ce qui correspond au traitement du langage et à l'analyse de flux de données temporels divers comme des électro-cardiogrammes. Le tout, sans plus de détails techniques !

  Tandis que les chipsets TrueNorth d'IBM ne gèrent pas l'apprentissage et ne font qu'exécuter les modèles neuronaux déjà entraînés, ici, le processeur est capable d'apprentissage et dans les modes supervisés, non supervisés et par renforcement. Ce processeur comprend deux chipsets, l'un qui a l'air d'avoir une architecture de CPU (en haut) et l'autre qui ressemble bien à un réseau neuromorphique (en bas).

[243]Relatées dans La puce Loihi d'Intel aussi performante qu'un cerveau de souris en 2019 de Mark Hachman en juin 2018.

Quittons Intel et voyons qui d'autre s'active dans les chipsets pour serveurs :

- **Bitmain** (2013, Chine, $450M), une unicorn chinoise bien connue, à l'origine des chipsets de mining de Bitcoins, les AntMiner 9 qui dominent le marché. Elle a aussi développé un chipset d'IA, le Sophon BM1680 (*ci-contre*), qui sert à la fois à l'entraînement et à l'inférence de réseaux de neurones, voisin des TPU de Google. Il est destiné aux applications de traitement de l'image comme du langage, et notamment aux systèmes de vidéo surveillance dont la Chine rafole comme l'illustre le positionnement de nombreuses startups du pays.
- **Graphcore** (2016, UK, $110M) qui conçoit son Intelligence processuing Units (IPU), un chipset adapté à l'exécution d'applications de deep learning côté entraînement et inférence qui comprendrait 1000 cœurs.
- **Cerebras Systems** (2016, USA, $112M) est une étonnante startup ayant déjà atteint une valorisation de $860M, lui donnant un statut envié de pré-licorne mais qui fonctionne en mode silencieux ("*stealth mode*"). On sait juste qu'elle conçoit un ASIC pour de l'entraînement de réseaux de neurones. La startup a été créé par des anciens de SeaMicro, une startup constructeur de serveurs à basse consommation acquise par AMD en 2012 pour $357M, complétée récemment par un dirigeant d'Intel, Dhiraj Mallick.
- **Groq** (2017, USA, $10M), une startup créée par des anciens de Google qui avaient participé à la conception de leurs TPU. Leur chipset pour serveur est censé générer 400 Tflops avec 8 Tflops/s par Watt.
- **Gyrfalcon Technology Inc** ou GTI (2017, USA) est sorti du bois en septembre 2017 avec deux chipsets d'inférences ASIC à basse consommation, l'un pour les serveurs et l'autre pour les objets connectés.La version serveur (Lightspeeur 2803 AI Accelerator) est intégrée dans des cartes à 16 composants[244].
- **Wave Computing** (2010, USA) qui développe ses Dataflow Processing Units avec 16 000 cœurs produits en ASIC chez TSMC en 16 nm, dédiés à l'entraînement de réseaux de neurones. Ces DPU sont assemblés dans des serveurs par paquets de 16, donnant 128 000 cœurs. Ils n'utilisent par contre que de la DRAM, bien moins performante que la mémoire HBM des GPU Nvidia et autres chipsets plus spécialisés[245]. La startup avait fait l'acquisition de l'activité MIPS de l'Anglais Imagination Technologies.

2.9 PetaOPS /sec	5.8 PetaOPS /sec	8.7 PetaOPS /sec	11.6 PetaOPS /sec
16 DPUs	32 DPUs	48 DPUs	64 DPUs
128GB High Speed Memory	256GB High Speed Memory	384GB High Speed Memory	512GB High Speed Memory
16TB SSD Storage	32TB SSD Storage	48TB SSD Storage	64TB SSD Storage
2TB Bulk Storage	4TB Bulk Storage	6TB Bulk Storage	8TB Bulk Storage

Up to Four 3U Wave Computers Per a Single Data Center Node

- **ThinkForce Electronic Technology** (2017, Chine, $68M) qui développe des chipsets de deep learning pour serveurs, également basés sur des architectures multicœurs.
- **Habana Labs** (2016, Israël, $75M) commercialise une carte PCIe pour serveur comprenant leur processeur Goya HL-1000 qui peut traiter 15 000 images/second avec seulement 100 W. Le système supporte TensorFlow et le format d'échange ONNX.

---

[244] Voir AI Accelerator Gyrfalcon Soars Post Stealth de Kevin Fogarty, novembre 2018.

[245] Source : AI processing requirements reveal weaknesses in current methods de Tom Simon, juillet 2018.

- **SambaNova Systems** (USA, $56M) développe aussi un chipset d'inférences pour serveurs qui cherche à dépasser la performance des Volta de Nvidia et cible le traitemet de données. La start-tup a aussi obtenu $8M de financement de la DARPA.
- **Baidu** qui présentait en juillet 2018 son chipset serveur **Kunlun** réalisé en 14 nm chez Samsung qui semble être généraliste, étant adapté aussi bien à la reconnaissance d'images qu'au traitement du langage. C'est visblement dans un premier temps un FPGA. Il a une bande passante mémoire convenable de 512 Go/s et délivre 260 teraops pour une consommation de 100W.
- **Alibaba** et **Tencent** ont aussi leurs chipsets serveurs de machine learning pour serveurs, aussi réalisés en FPGA.

Nous allons maintenant passer aux processeurs adaptés à l'exécution d'applications de deep learning dans l'embarqué, à la fois dans les smartphones et dans les objets connectés en général. On appelle cela l'Edge AI, ou l'IA appliquée aux objets. L'offre y est bien plus abondante que sur serveurs.

Sur smartphone, le premier des chipsets d'IA en date est le Kirin 970 de **HiSilicon**, la filiale de semiconducteurs du chinois Huawei. Présentée à l'IFA 2017, il s'agissait d'un chipset mobile gravé en 10 nm par TSMC et comprenant 5,5 milliards de transistors.

Il comprend un NPU (Neural Processing Units) faite de multiplicateur de matrices 3x3 qui est dédiée au traitement d'applications de deep learning comme la reconnaissance de la parole ou d'images qu'ils appellent une Neural Processing Unit (NPU).

On l'a retrouvé dans de nombreux smartphones lancés depuis comme les Huawei Mate 10, Pmate 20 et Honor 10. Le tout est complété de 8 cœurs CPU et 12 cœurs GPU MALI (design de GPU d'origine ARM). Le NPU peut traiter 1,92 TFlops en calculs flottants FP16. Dans le Kirin 980, la puissance est doublée à 4 Tflops/s. Le NPU supporte Tensorflow, Tensorflow Lite et Caffe/Caffe2. Ce chipset a été suivi en septembre 2018 du Kirin 980 qui double la puissance côté NPU, et est gravé en technologie 7 nm.

Les Kirin 970 et 980 utilisent une conception de circuit provenant de **Cambricon Technology** (2016, Chine, $200M[246]). HiSilicon n'a pas utilisé tel que, un bloc de processeur neuromorphique de Cambricon Technology. Ils ont travaillé ensemble pour le personnaliser et l'intégrer dans le Kirin 970 puis dans le Kirin 980 et notamment pour l'adapter au processus de fabrication du chipset qui est en intégration à 10 nm puis en 7 nm, fabriqué par TSMC à Taïwan.

Une semaine après l'annonce du Kirin 970 par Huawei, **Apple** lançait sa nouvelle salve d'iPhones 8 et X. Ceux-ci intègrent aussi une fonction neuromorphique sous la fome d'un coprocesseur dénommé A11 Bionic Neural Engine. Il tourne à 900 MHz mais rien n'a encore filtré sur ses capacités techniques précises. On sait sans surprise qu'il est exploité par SIRI et par les fonctions de reconnaissance d'images comme le login exploitant une vue 3D du visage.

[246] Ces $200 ont été levés auprès de SDIC, un investisseur public chinois qui ressemble à notre Bpifrance. Fin 2017, Cambricon annonçait lancer d'ici 2019 la fabrication de son propre chipset en 16 nm chez TSMC, les MLU100 et MLU200 (Machine Learning Units) dédiés aux serveurs.

Le chipset A12 équipant les iPhone XR lancés en septembre 2018 double la puissance du NPU à environ 3-4 TeraOps, sans que l'on ne sache ce qu'il contient.

Au CES2018, le Chinois **Rockchip** sortait son premier processeur embarqué RK3399Pro intégrant un NPU, atteignant 2,4 Tflops, au niveau du Kirin 970 de HiSilicon. Il comprend sinon huit cœurs ARM en architecture big.LITTLE Cortex-A72 et Cortex-A53 plus un GPU Mali-T860. Le NPU sert notamment à la reconnaissance d'images et de la parole. Il est supporté par les frameworks de machine learning OpenVX (open source, pour la vision), TensorFlow (généraliste) et Caffe (également généraliste).

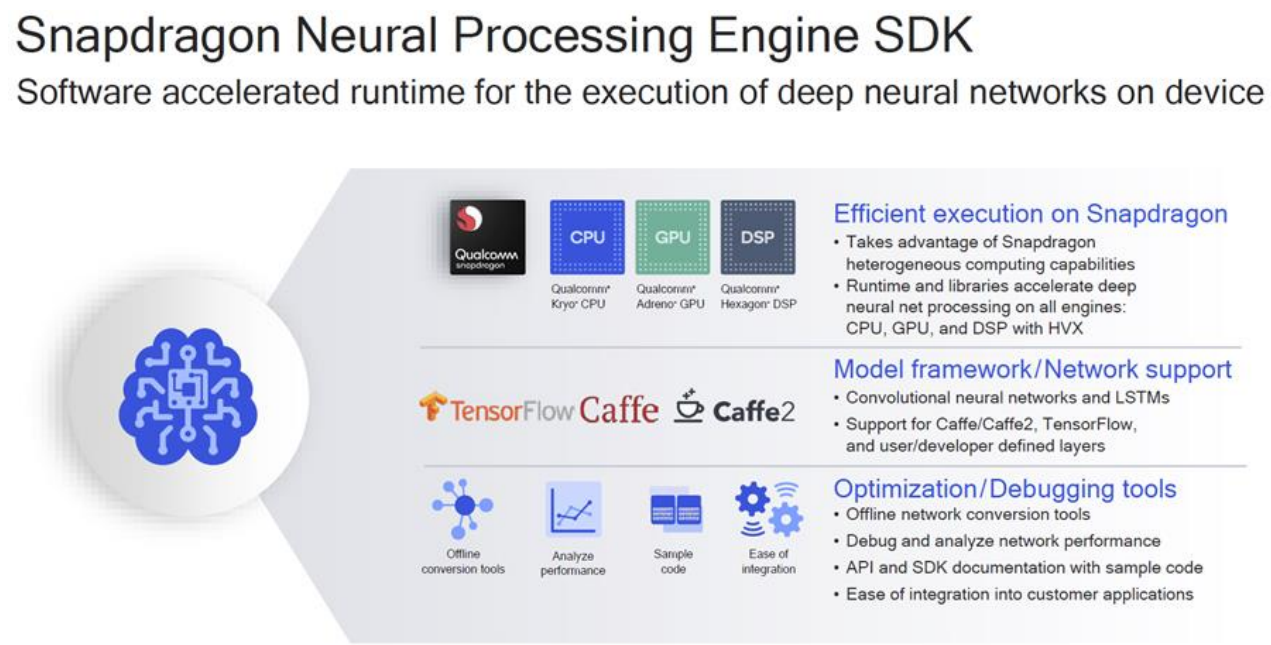
Le Taïwanais **MediaTek** faisait de même en annonçant sa plateforme NeuroPilot, une "AI processing unit" (APU) associée à un SDK NeuroPilot qui supporte les habituels outils de l'IA tels que TensorFlow, Caffe et Amazon MXNet.

Les marchés visés sont les smartphones et l'automobile. L'annonce ne précisait pas les fonctions mathématiques mises en œuvre dans leur APU, ce qui est bien dommage mais risque d'être courant.

Chez **Qualcomm**, l'approche vis-à-vis de l'IA est très différente[247]. Elle est gros basée sur la création du Snapdragon Neural Processing Engine SDK qui supporte d'un côté les principaux frameworks de deep learning du marché (Tensorflow, Caffe, Caffe2, ONNX et les API Android Neural Networks) et de l'autre, qui exploite les différentes composantes des chipsets Snapdragon, les cœurs Kryo, le GPU maison Adreno et les DSP Hexagon qui contiennent des unités de traitement de vecteurs.

On n'a donc pas à proprement parler de NPU. Le DSP Hexagon comprend toutefois des unités de manipulation de vecteurs qui optimisent le fonctionnement des réseaux de neurones, une archirecture voisine de ce que l'on trouve dans les processeurs serveur Intel Xeon Phi. La bibliothèque Hexagon Neural Network permet d'exécuter des logiciels de deep learning directement sur les processeurs vectoriels Hexagon, notamment pour des réseaux convolutionnels.

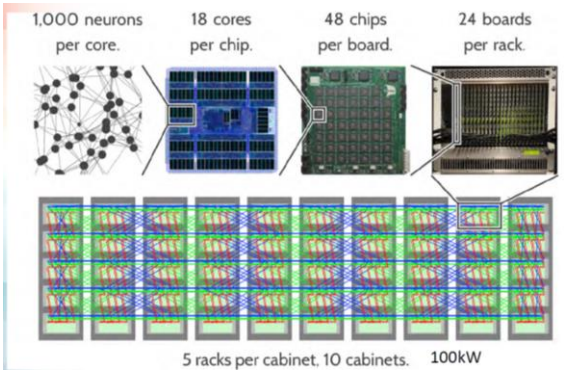
Qualcomm s'appuie notamment sur les compétences des équipes de **Scyfer** (2013, Pays-Bas), une startup issue de l'Université d'Amsterdam acquise en 2017 et spécialisée dans les développements logiciels en machine learning. Ils font aussi appel à **Brain Corp** (2009, USA, $125M) dans lequel Qualcomm Ventures a investi et pour de la R&D externalisée dans la vision artificielle.



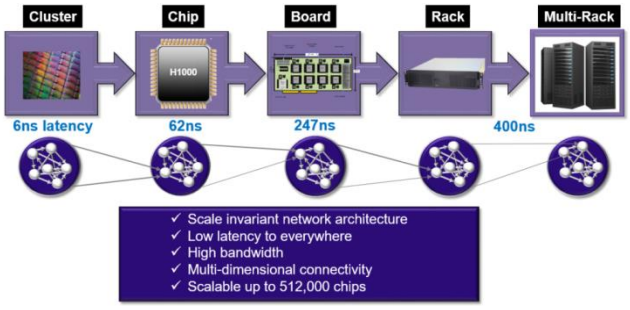
[247] Voir Making on-device AI ubiquitous, mars 2018 (19 slides).

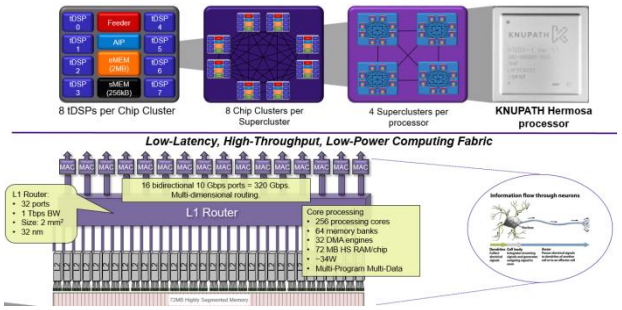
Dans l'embarqué, nous avons aussi une belle brochette d'offres disponibles ou en cours de développement :

- **Mythic** (2012, USA, $55M) conçoit des chipsets pour micro-ordinateurs qui sont optimisés du côté de la gestion de la mémoire et avec une interface PCIe. Ils utiliseraient une méthode associant numérique et anologique pour accélérer les inférences dans les réseaux de neurones. Le chipset sera notamment supporté par TensorFlow[248].
- **Prophesee** (2013, France, $37M) est l'ex Chronocam, chez qui Intel est le plus gros investisseur. Leur chipset est en fait un capteur vidéo qui intègre un réseau de neurones permettant l'interprétation immédiate des images.
- **SpiNNaker** a été créé dans le cadre du projet européen Human Brain Project par Steve Furber (Université de Manchester, UK). Il vise à simuler le fonctionnement d'un milliard de neurones. Il s'appuie sur une architecture matérielle avec 18 cœurs 32 bits ARM de 1000 neurones par chipset. On est plus dans l'architecture massivement parallèle avec des milliers de processeurs de ce type que dans les processeurs véritablement synaptiques.



  L'architecture est suffisamment souple pour exécuter différents types de réseaux de neurones y compris des neurones à impulsions.
- **KnuEdge** (2007, $100M) planche sur un chipset Knupath qui est basé sur la technologie LambaFabric qui permet l'alignement en parallèle de 512 000 unités assemblées dans des chipsets de 256 cœurs. L'offre comprend les chipsets KnuVerse dédié à la reconnaissance de la parole pour l'authentification ainsi que les services en cloud Knurld.io permettant d'intégrer l'authentification vocale dans une application.





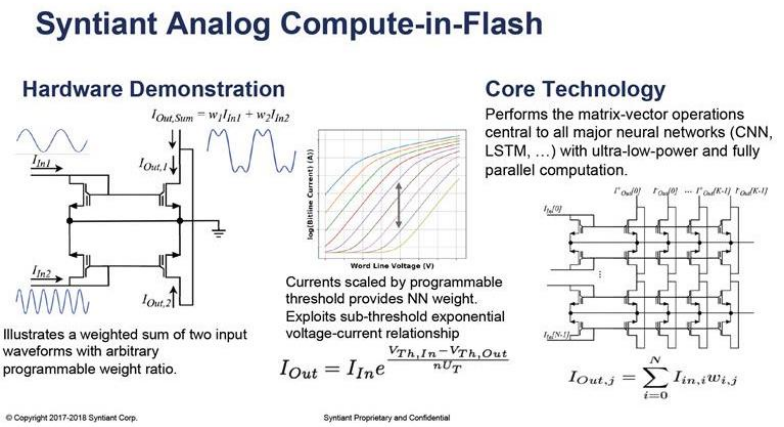
- **Neurocore** est un projet Stanford, aussi associé au HBP, un chipset intégrant 65536 neurones et fonctionnant à très basse consommation.
- **AnotherBrain** (2017, France, $10M) est une startup lancée par Bruno Maisonnier, le fondateur d'Aldebaran Robotics. L'architecture de son chipset n'est pas documentée à ce stade. Le concept serait très différent de tous les autres chipsets neuromorphiques, avec comme avantage, un processus d'entraînement plus rapide et nécessitant moins de données. On attend maintenant de juger sur pièces !
- **Vathys** (2015, USA, $120K) développe un chipset censé être 10 fois plus rapide que les concurrents en optimisant le mouvement des données. Ils en sont pour l'instant à l'état du concept[249].

[248] Voir Mythic nets $40M to create a new breed of efficient AI-focused hardware de Matthew Lynley, mars 2018.

[249] Voir les détails dans Vathys Petascale Deep Learning on a (Single) Chip, 2017 (28 slides).

- **Horizon Robotics** (2015, Chine, $100M, provenant notamment d'Intel Capital), est une startup spécialisée dans les composants pour la robotique et la conduite autonome. Leur chipset Sunrise 1.0 sert à la reconnaissance de visages dans des vidéos, supportant une entrée en Full HD et la suivi simultané de 200 objets avec une consommation de seulement 1,5W. Un autre chipset, le Journey 1.0 suit les trajectoires de huit catégories de cibles comme les piétons, les vélo, les voitures et la signalisation routière. Cela cible les systèmes d'ADAS, pour la conduite assistée ou autonome.

- **Syntiant** (2017, USA, $5,1M) développe le chipset Neural Decision Processors (NDP) qui intègre une sorte de mémoire Flash analogique à côté de ses fonctions d'inférence de deep learning organisées sous forme de multiplicateur de matrices par des vecteurs, le tout fonctionnant à basse consommation grâce à des synapses avec une basse précision de 3 à 5 bits.



  Le chipset génère 20 teraops par Watt. Il vise bien entendu les applications mobiles en tant que coprocesseur[250].

- **Novumind** (2015, USA, $15,2M) développe ses chipsets NovuTensor, dédiés aux inférences de réseaux de neurones convolutionnels pour la reconnaissance d'images. Ils génèrent 15 Tflops/s pour 5 Watts et 3 Tflops par Watt, ce qui semble être maintenant la norme. Le NovuTensor est disponible sous la forme de composant ou dans une petite carte PCIe qui peut s'installer dans un serveur. L'ensemble était en bêta en date de septembre 2018.

- **DeePhi Tech** (2016, Chine, $40M) prévoit de sortir des chipsets neuromorphiques en 2018, l'un pour le cloud et l'autre pour l'embarqué et spécialisés dans la reconnaissance d'images. Leur DPU (Deep Learning Processor Unit) est un FPGA complété par le SDK Deep Neural Network Development Kit (DNNDK). Cette spin-off issue d'un partenariat entre l'Université de Tsinghua et celle de Stanford vise le marché des caméras de surveillance et des robots. Elle a été acquise par l'Américain Xilinx en juillet 2018 qui propose par ailleurs son DNN Processor pour les FPGA Xilinx avec ses outils logiciels associés[251]. Xilink est le numéro un mondial des FPGA, devant Intel.

- **GrAI Matter Labs** (2016, France, $15M) conçoit un processeur sur une architecture originale, a base de réseaux neuronaux, numérique mais asynchrone, utilisant des « spiking neurons » (neurones à impulsion). Ils ambitionnent d'intégrer un million de neurones sur un centimètre carré et consommant 1W, programmable en Python. Ils ont obtenu $1M de financement par la DARPA pour un démonstrateur FPGA qui tourne déjà. C'est rare pour une startup française.

- **Kalray** (2008, France, $39,7) est une startup spin-off du CEA qui développe des processeurs « multicore », à 256 cœurs, dédiés au marché de l'automobile et aux applications d'IA associées. La startup annonçait en mai 2018 l'entrée à son capital d'Alliance Ventures, le fonds de capital-risque de l'Alliance Renault-Nissan-Mitsubishi, et de Definvest, le fonds géré par Bpifrance pour le compte du Ministère des Armées.

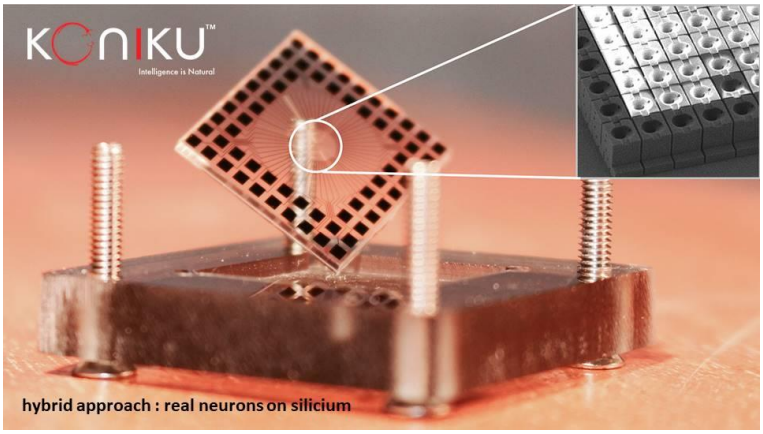
---

[250] Le schema provident de la presentation Analog Computers for Deep Machine Learning de Jeremy Holleman, le CTO de Syntiant (19 slides).

[251] Voir Xilinx DNN Processor, 2018 (20 slides).

- Le **CogniMem** (2011, France/USA) propose le CM1K est un chipset ASIC intégrant un réseau de 1024 neurones stockant chacun 256 octets qui sert aux applications de reconnaissance des formes. Ne coutant que $94, il est notamment utilisé dans la **BrainCard**, issue de la start-up franco-américaine, **General Vision** qui commercialise des « blocs d'IP » pour créer des processeurs neuromorphiques, avec ses NeuroMem. Cette technologie est aussi intégrée dans les processeurs d'objets connectés Curie d'Intel (avec 128 neurones, mais abandonnés par ce dernier en juillet 2017). L'ensemble sert principalement aux applications de vision artificielle dans les systèmes embarqués.

- Enfin, terminons ce panorama bigaré avec **Koniku** (2014, USA, $1,65M), une startup qui développe des neurones hybrides en silicium et biologiques. La société californienne se positionne aussi sur la reconnaissance d'images. Ils visent les marchés militaires, de la sécurité et de l'agriculture. Mais ils ne donnent pas beaucoup de signes de vie.



Pour être complet, il faudrait aussi citer quelques autres startups : **ThinCI** (2010, USA, $65M) et ses Graph Streaming Processor ASIC pour de la vision artificielle, **TensTorrent** (2016, Canada, $500K) également avec des ASIC pour la vision, **Kneron** (2016, Chine, $33M) avec ses ASIC pour la vision et **Hailo** (2017, Israël, $12,5M) avec ses ASIC généralistes.

Les concepteurs de blocs fonctionnels pour processeurs embarqués qui les commercialisent sous forme de propriété intellectuelle se sont aussi mis à l'IA depuis 2016. Les « blocs d'IP » sont des blocs fonctionnels de chipsets qui sont vendus aux concepteurs de chipsets sous forme de propriété intellectuelle (« blocs de propriété intellectuelle »). Ces blocs fonctionnels sont souvent dédiés soit au traitement de l'image soit à celui du langage en parfois du bruit.

- **Tensilica** (1997, USA, $44M) avec ses Vision C5 surtout destinés aux systèmes de vision artificielle comme les caméras de surveillance. En septembre 2018, Cadence, qui possède Tensilica, annonçait les blocs d'IP Tensilica DNA 100 également adaptés à la vision artificielle, permettant de réaliser des inférences sur plus de 2500 images par seconde[252].

- **eSilicon** (1999, USA, $98M) propose son bloc d'IP neuASIC pour chipsets en technologie ASIC 7nm pour exécuter des applications de machine learning (détails).

- **Teradeep** (2014, USA) propose des blocs fonctionnels pas bien documentés, les Pure RTL.

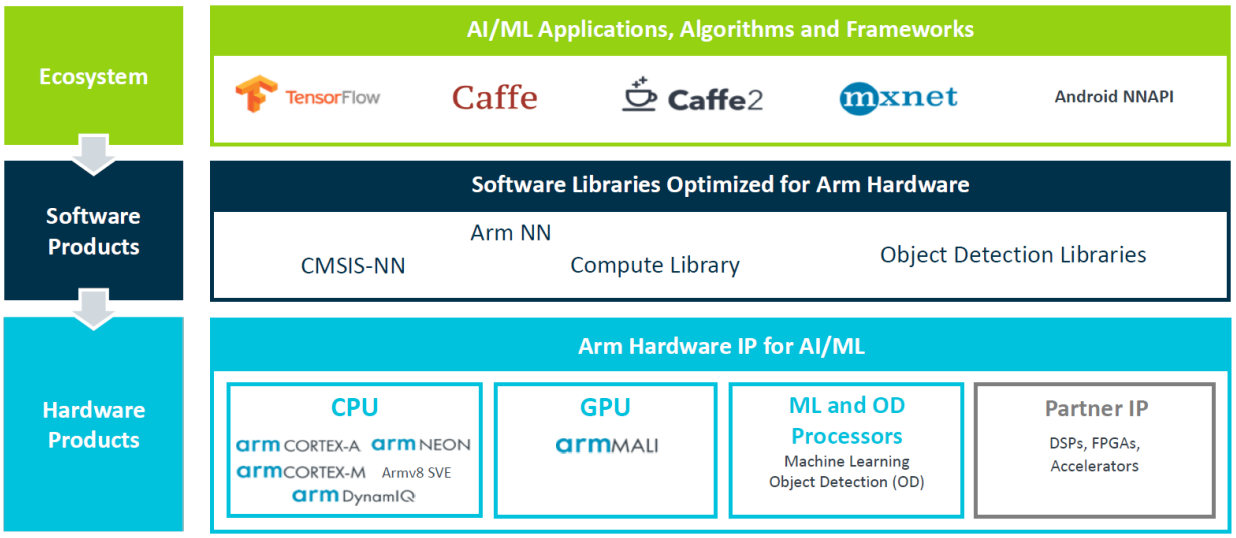
- **Arm** est un peu à la traîne sur l'IA et déploie une stratégie voisine de celle de Qualcomm avec en premier lieu une offre logicielle dans la plateforme du projet Trillium annoncée en février 2018. Comme il se doit, l'ensemble supporte les principaux frameworks de deep learning leaders du marché (TensorFlow, Caffee2, Mxnet, Android NNAPI) avec diverses bibliothèques Arm qui les relient aux blocs d'IP arm, notamment les Cortex-M utilisés dans de nombreux micro-contrôleurs, les GPU Mali et un nouveau bloc d'IP dédié à la détection d'objets dans des réseaux convolutionnels[253] avec 4 teraops/s visant 3 teraops par Watt consommé[254] ce qui ne constitue pas une avancée particulière.

---

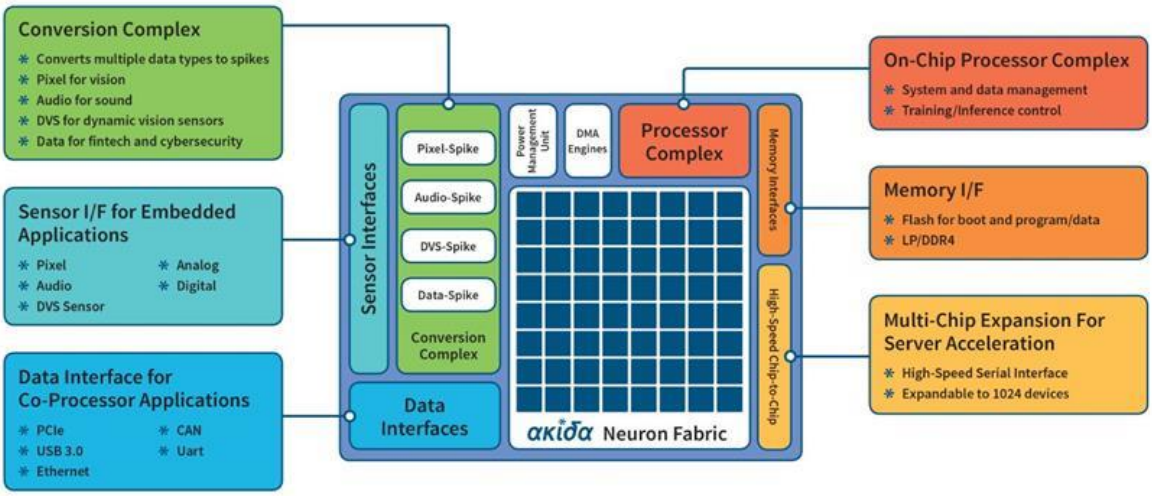
[252] Voir Cadence Announces The Tensilica DNA 100 IP: Bigger Artificial Intelligence, septembre 2018.

[253] Voir Deep Learning on Arm Cortex-M Microcontrollers de Rod Crawford, Arm (28 slides)

[254] Ses fonctionnalités sont détaillées dans Hot Chips 2018: Arm's Machine Learning Core Live Blog, avril 2018.



- **Leapmind** (2012, Japon, $13,5M) qui créé des blocs fonctionnels pour réseaux de neurones à basse consommation.
- **Brainchip** (2006, Australie, $27,8M) développe des modules de processeurs à neurones par impulsion (spiking neuron adaptive processor ou SNAP) qui sont licenciés à des concepteurs de chipsets. Ils développent aussi les briques logicielles qui exploitent l'architecture. Les *spiking neurons* sont plutôt adaptés au traitement du langage et des réseaux récurrents. Leurs chipsets Akida comprennent 1,2 million de neurones et 10 milliards de synapses. D'après la littérature de la startup, ils permettent un apprentissage « non supervisé » à la fois pour le traitement du langage et de l'image[255].



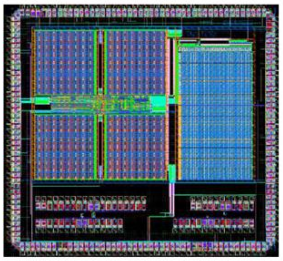
[255] Source du schéma : https://www.brainchipinc.com/products/akida-neuromorphic-system-on-chip. Et la présentation du lancement d'Akida : Brainchip Akida Launch Presentation, 2018 (27 slides).

- Le **CEA LETI** a commencé à travailler sur les chipsets neuromorphiques N2-D2 produits en technologie FD-SOI 28 nm et dédiés aux systèmes embarqués et à la reconnaissance d'images. Il s'est alors impliqué dans la conception de la puce basse consommation Dynapse-SL du projet H2020 européen Neuram3 impliquant notamment la Suisse et IBM[256].

DYNAPS-SL (INI-ZURICH)

	Dynapse-SL	IBM True North
Technology	28 nm FDSOI	28nm CMOS
Supply Voltage	1 V	0.7V
Neuron Type	Analog	Digital
Neurons per core	256	256
Core Area	0.36 mm²	0.094 mm²
Computation	Parallel processing	Time multiplexing
Fan In/Out	2k/8k	256/256
Synaptic Operation per Second per Watt	300 GSOPS/W*1	46 GSOPS/W
Energy per synaptic event	<2 pJ*2	10 pJ
Energy per spike	<0.375 nJ*3	3.9 nJ

* 1 At 100Hz mean firing rate, by appending 4 local-core destinations per spike, 400 k events will be broadcast to 4 cores with 25% connectivity per event: 400 k x 1 k x 25% / 300 µ W = 300 GSOPS/W
* 2 In case of 25% match in each core, energy per synaptic event = energy per broadcast / (256*25%) =120pJ/64 = 2 pJ
* 3 Energy per spike = total power consumption / spikes numbers = 300 uW/800 k = 0.375 nJ



- **STMicroelectronics** investi aussi le domaine de l'IA dans l'embarqué avec son kit logiciel STM32CubeMX.ai servant à compiler du code de réseau de neurones entraîné et à l'exécuter sur un microcontrôleur classique STM32 à basse consommation. Il supporte les principaux framework open source du marché comme TensorFlow. STMicroelectronics présentait aussi au CES 2018 un prototype de chipset réalisé en 28 nano FD-SOI pour l'exécution directe de réseaux de neurones, comprenant 8 accélérateurs de réseaux convolutionnels, 2 DSP et un cœur ARM Cortex M4. Il ne semble pas encore disponible.

Memristors

Les memristors ont été conceptualisés en 1971 par le sino-américain **Leon Chua**[257]. Ce sont des composants électroniques capables de mémoriser un état en faisant varier une résistance électrique par l'application d'une tension, un peu comme les cristaux liquides bistables qui servent dans les liseuses électroniques.

La valeur modifiable de la résistance permet de stocker de l'information et de manière non volatile. C'est particulièrement utile pour créer des réseaux de neurones avec des poids de synapses gérés dans ces résistances.

Ces poids peuvent être lus et modifiés rapidement sans avoir besoin d'accéder à une mémoire externe, qui bien est lente par rapport à la vitesse de fonctionnement du processeur, en particulier dans les phases laborieuses d'entraînement du deep learning. Dans certains cas de figure, cela permet même de réaliser un entraînement temps réel incrémental de réseau de neurones. C'est un type d'architecture de chipsets envisagée depuis des années mais difficile à réaliser[258].

Les memristors peuvent aussi servir à la partie logique des composants et remplacer des transistors classiques ou bien être intégrés au côté de transistors actifs classiques dans des unités de traitement[259]. Dans le premier cas, cela pourrait aboutir à une nouvelle génération de circuits FPGA dont la logique est dynamiquement programmable. Le second cas n'est pas encore opérationnel.

---

[256] Quelle différence entre ces deux projets? N2-D2 est un chipset neuromorphique classique qui intègre la logique de traitements. Dynapse-SL est un chipset qui intègre la logique et la mémoire, avec des neurones analogiques, permettant d'améliorer les performances et de réduire encore plus la consommation d'énergie. Le CEA-LETI bosse aussi sur des chipsets avec des neurones à impulsion, dans le cadre du projet Spirit et exploitant de la mémoire OxRAM. Voir Neuromorphic and Deep Learning Technologies at CEA, 2017 (57 slides), NeuRAM3: NEUral computing aRchitectures in Advanced Monolithic 3D-VLSI nano-technologies, 2017 (7 slides) et OxRAM Memories A Disruptive Technology For Disruptive Designs, 2017 (30 slides).

[257] Dans Memristor-The Missing Circuit Element de Leon Chua, 1971 (13 pages).

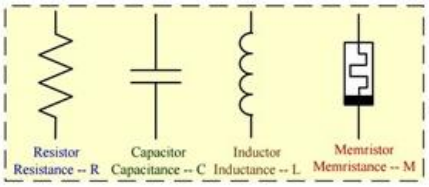
[258] Voir The Search for Alternative Computational Paradigms de Naresh Shanbhag, 2008 (11 pages) qui décrit diverses architectures rapprochant calcul et mémoire. Ainsi que Cognitive Computing at the Limits du même auteur, 2017 (19 slides).

[259] C'est très bien expliqué dans Memristor: From Basics to Deployment de Saraju Mohanty, publié en 2013 (13 pages) et dans Hybrid CMOS/Memristor Circuits 2010 (4 pages).

Par contre, les memristors sont difficiles à mettre au point car ils ne restent pas stables longtemps et leur temps de commutation est parfois trop long. Ils sont aussi difficiles à miniaturiser au même niveau que les mémoires volatiles et non volatiles actuelles.

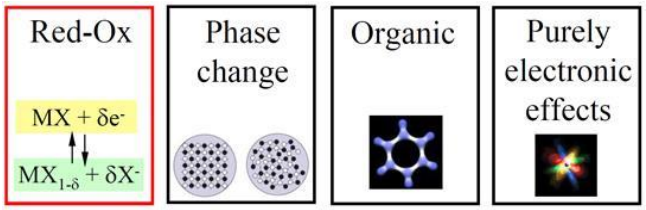
Les memristors sont produits avec des procédés de fabrication voisins de ceux du CMOS, en ajoutant une bonne douzaine d'étapes dans la production avec des procédés spécifiques de dépôt sous vide de couches minces de matériaux semi-conducteurs.

Les fabs investies sur le sujet sont notamment celles de Global Foundries, TSMC et Samsung. Placer la mémoire à proximité des neurones permet aussi d'économiser de l'énergie et de se rapprocher un tant soit peu de l'efficacité énergétique extraordinaire du cerveau humain, qui ne consomme que 20 Watts.



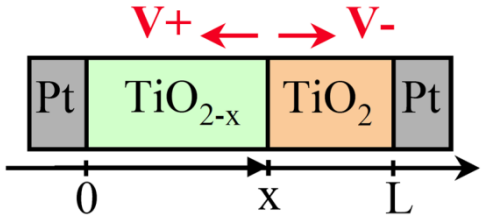
En plus des réseaux de neurones, les memristors ont évidemment comme domaine d'application les mémoires non volatiles pour créer des SSD. Le défi est de descendre en-dessous de 15 nm pour la taille de chaque cellule mémoire et de supporter un grand nombre de cycles d'écriture et de lecture.

Il existe différents types de memristors qui ont chacun leurs avantages et inconvénients. La recherche progresse depuis près de 20 ans dans le domaine, mais on n'est pas encore au stade de l'industrialisation.

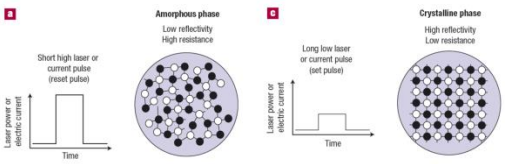


On compte ainsi :

- Les **Red-Ox**[260], des résistances variables par migration d'ions d'oxyde de titane, de tantale, d'argent ou de silicium[261], par procédé de réduction-oxydation. Les ReRAM annoncées par HP en 2013 en font partie. Depuis, HP travaille sur leur fiabilité, leur vitesse de commutation et leur endurance[262].



- Les memristors à **changement de phase** ou PCM (Phase-Change Memory) qui alterne phase amorphe et phase cristalline pour des composites comme du verre de chalcogénure GST (germanium-antimoine-tellure)[263].



  Les chipsets pour SSD 3D Xpoint d'Intel utilisent une variante de ce type de memristors. On y range les PRAM et les PCRAM.

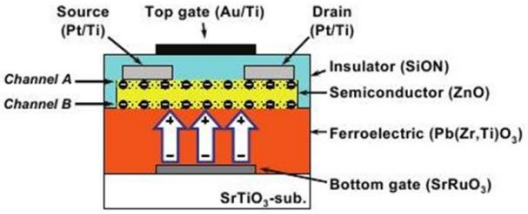
[260] Voir Memristive devices for computing de Joshua Yang, Dmitri Strukov et Duncan Stewart, 2013 (12 pages).

[261] Weebit Nano est une startup lancée en 2015 en Australie qui développe de la ReRAM à base d'oxyde de silicium. Le SiOx est utilisé dans les transistors. La startup a démontré en juin 2018 une première puce de 1 mbits en technologie 40 nm. La démonstration avait lieu au CEA-LETI à Grenoble, leur partenaire.

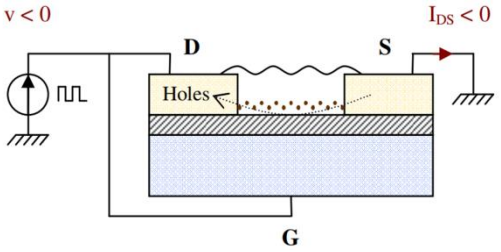
[262] Les schémas sont issus de la présentation Neuromorphic computing - Memristors de Julie Grollier, du CNRS / Thalès TRT, 2018 (80 slides). Voir aussi dans son blog : http://julie.grollier.free.fr/memristors.htm. Le Coréen SK Hynix met au point de son côté une ReRAM en technologie 23 nm à base se silicium dopé à l'arsenic.

[263] Cette technologie a été créée par une chercheuse Turco-Américaine, Duygu Kuzum.

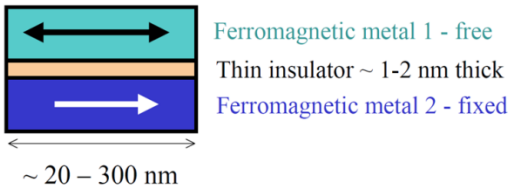
- Les memristors **ferroélectriques** qui exploitent un changement de polarité par champ électrique (source du schéma). La technique exploite d'or, du plomb, du zirconium et du ruthenium qui ne sont pas des éléments très abondants.



- Les memristors **organiques**[264], avec des polymères divers comme les NOMFET (Nanoparticle Organic Memory Field-Effect Transistor) avec de l'or et du pentacène, un hydrocarbure à base d'hydrogène et de carbone, des nanotubes de carbone (CNT-FET) et des polyanilines (PANI)[265].



- Les memristors à base de **spintronique** qui sont utilisables dans des MRAM (Magnetic RAM) avec la modification d'un état magnétique qui affecte une résistance, via un isolant entre deux métaux ferro-magnétiques.



L'un des problèmes à résoudre est la durabilité des memristors, notamment pour ceux qui exploitent des techniques qui déplacent des atomes. Ceux-ci claquent au bout d'un million de cycles du fait de failles crisallines. C'est moins le cas avec les memristors à changements de phase. On peut aussi éviter les mouvements d'atomes en jouant sur des jonctions à effet tunnel et à mouvements d'électrons.

Dans le cas de memristors utilisés pour des poids synaptiques, il faut aussi réduire le bruit et les erreurs qui génèrent des poids imprécis. L'autre solution demanderait d'exploiter des réseaux de neurones à synapses binaires, avec seulement deux valeurs comme dans les perceptrons d'origine. Yoshua Bengio de l'Université de Montréal travaille là-dessus, avec les **Quantized Neural Networks**[266] (QNN) qui auraient des performances voisines vis-à-vis des réseaux de neurones classiques, comme pour la reconnaissance d'images dans le jeu de test ImageNet.

Des memristors ont été notamment développés dans le cadre des projets de recherche du programme **SyNAPSE** de la DARPA, financé entre 2008 et 2016. **HP** a été le premier à en prototyper en 2008, avec de l'oxyde de titane.

HP a même lancé un partenariat avec le fabricant de mémoires **Hynix**, mais le projet a été mis en veilleuse **en 2012**. Le taux de rebus serait trop élevé lors de la fabrication. C'est un paramètre clé pour pouvoir fabriquer des composants en quantité industrielle et à un prix de vente abordable. De plus, le nombre de cycles d'écriture semblait limité pour des raisons chimiques, dans le cycle de libération/captation d'oxygène pour les memristors en oxydes de titane.

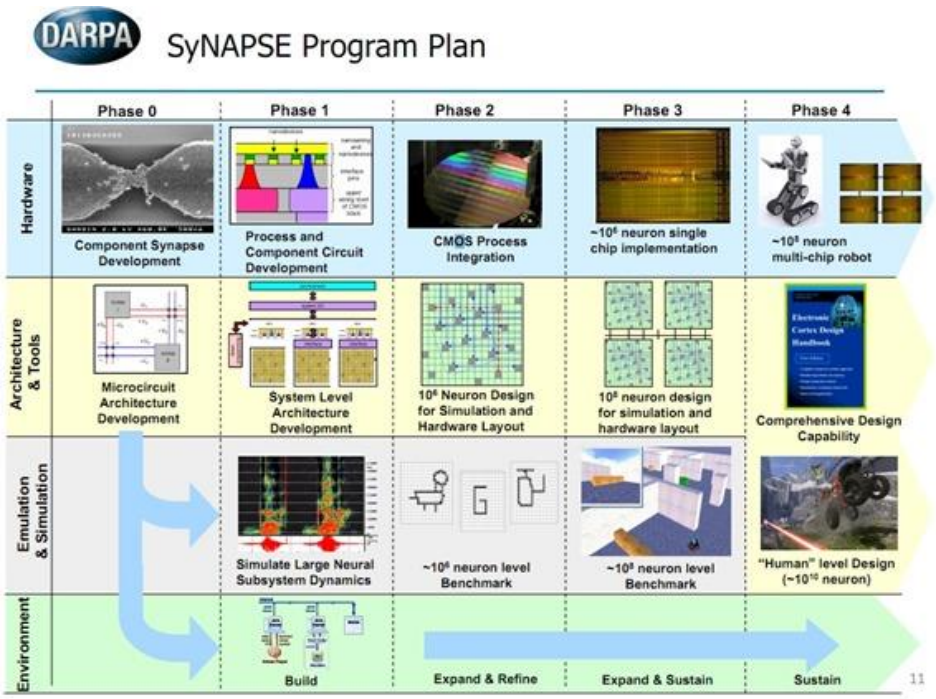
En octobre 2015, HP et **SanDisk** annoncaient un partenariat pour fabriquer des mémoires volatiles et non volatiles à base de memristors, censées être 1000 fois plus rapides et plus endurantes que les mémoires flash traditionnelles.

---

[264] Voir Organic memristors come of age de Ilia Valov et Michael Kozicki, 2017 (2 pages).

[265] Une équipe de chercheurs de l'Université de Singapour menée par Thirumalai Venky Venkatesan met au point des memristors capable de changer d'état en 50 ns et restant stable pendant 11 jours sans alimentation. Ils sont réalisés à base de ruthenium relies par des molécules organiques azotées conçues à l'Université de Yale et réalisées en Inde. Voir Organic Memristor Sets Records for Speed and Durability, octobre 2017.

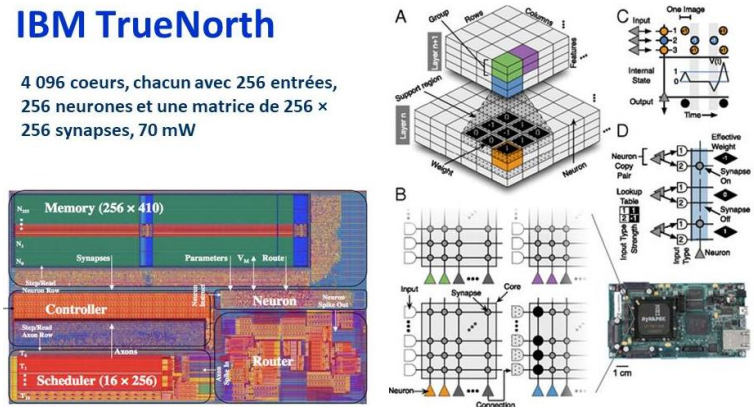
[266] Voir Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations, 2018 (30 pages).



Il nous faut ici évoquer le cas particulier d'**IBM** avec ses processeurs neuronaux TrueNorth, développé dans le cadre du programme SyNAPSE de la DARPA. Lancé en 2014, les chipsets TrueNorth sont capables de simuler un million de neurones artificiels, 256 millions de synapses reliant ces neurones et exécutant 46 milliards d'opérations synaptiques par secondes et par Watt consommé. Le tout avec 4096 cœurs fonctionnant par multiplexage temporel et 5,4 milliards de transistors, donc aux alentours d'une vingtaine de transistors par synapses.

TrueNorth utilise des neurones à impulsion (*spiking neurons*) qui imitent mieux le fonctionnement des neurones biologiques qui émettent des impulsions à intervalles réguliers. Ces neurones semblent plutôt adaptés au traitement du langage qu'à celui des images. Il faut noter que dans le cas de TrueNorth, ces neurones ne peuvent pas être exploités dans un processus d'apprentissage. Leur entraînement doit être réalisé sur d'autres machines.

Le chipset a été fabriqué par Samsung en technologie CMOS 28 nm et avec une couche d'isolation d'oxyde de silicium (SOI, avec des wafers issus du français **SOITEC**) permettant de diminuer la consommation électrique et d'accélérer les traitements. Le chipset fait plus de 4 cm^2 de surface et ne consomme que 70 mW, ce qui permet d'envisager d'empiler ces processeurs en couches.



C'est quelque chose de difficile à réaliser avec les processeurs CMOS habituels qui consomment beaucoup plus d'énergie au cm^2.

A titre de comparaison, un processeur Intel Core i7 de génération Coffee Lake réalisé en technologie 14 nm consomme entre 15 W et 130 W selon les modèles, avec 5 milliards de transistors et un GPU Nvidia V100 de 21 milliards de transistors consomme 300 W. Le but d'IBM est de construire un ordinateur doté de 10 milliards de neurones et 100 trillions de synapses, consommant 1 KW et tenant dans un volume de deux litres.

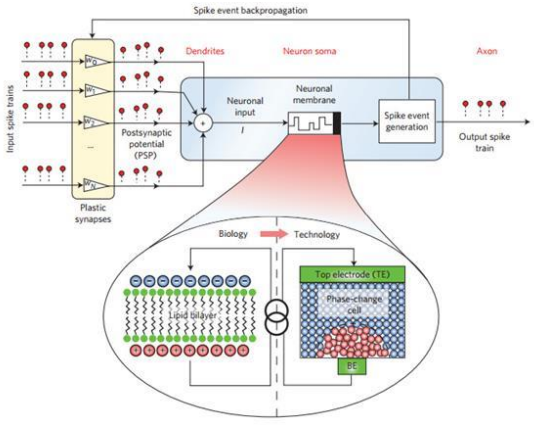
A titre de comparaison, un cerveau humain contient environ 86 milliards de neurones et ne consomme que 20 Watts ! Le biologique reste encore à ce stade une machine très efficace d'un point de vue énergétique ! Une étape intermédiaire a été annoncée au printemps 2017 : un ordinateur neuromorphique développé pour l'US Air Force et doté de 64 millions de neurones mais dont le domaine d'application n'a pas été précisé.

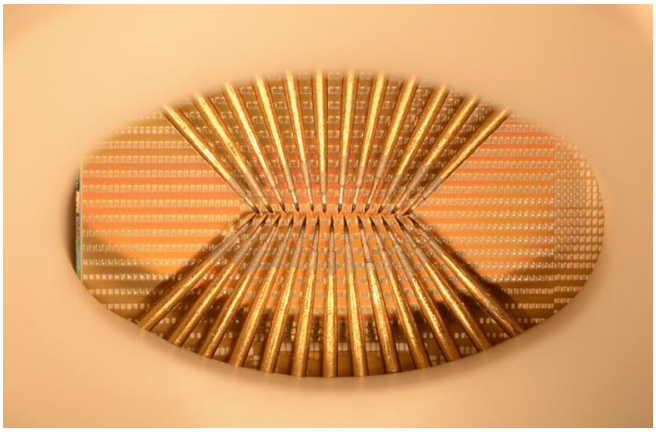
Mais cette première version de TrueNorth n'avait pas de mémoire intégrée. Il en va autrement d'une nouvelle mouture de chipset neuromorphique IBM testée en 2018 qui intègre des memristors PCM (Phase Change Memory)[267].

Cela a permis d'améliorer encore l'efficacité énergétique de l'ensemble, mais avec seulement 204 900 synapses pour commencer[268]. Les tests ont été réalisés sur un algorithme de reconnaissance d'écriture appliqué sur la base standard MNIST. Ce chipset réalisait 29 milliards d'opérations par seconde et par Watt consommé, un peu moins que le chipset TrueNorth d'origine.

D'autres laboratoires de recherche et industriels planchent aussi sur les memristors et les réseaux de neurones matériels :

- L'**ETH** de Zurich (le CNRS suisse) développe un memristor capable de stocker trois états à base de pérovskite (titanate de calcium) de 5 nm d'épaisseur[269]. Cela pourrait servir à gérer de la logique floue. Ils explorent aussi la piste des réseaux de neurones[270] à base de memristors à changement de phase en GST (germanium-antimoine-tellure).





[267] Voir Reliability enhancement of phase change memory for neuromorphic applications de SangBum Kim, IBM, 2017 (21 slides).

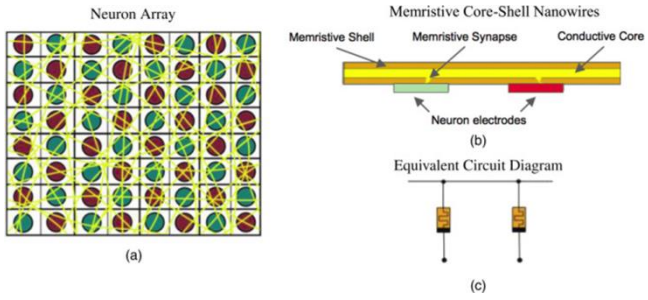
[268] AI could get 100 times more energy-efficient with IBM's new artificial synapses, juin 2018 et Equivalent-accuracy accelerated neural-network training using analogue memory, 2018. IBM communique sur le fait qu'ils utilisent une mémoire « analogique ». C'est en effet le cas pour les informations stockées dans les memristors.

[269] Voir Swiss researchers have created a memristor with three stable resistive states, 2015.

[270] Voir https://arstechnica.com/gadgets/2016/08/ibm-phase-change-neurons/.

- Des chercheurs de l'Université Technologique du Michigan **annoncaient début 2016** avoir créé des memristors à base de bisulfite de molybdène qui ont un comportement plus linéaire[271].
- Des chercheurs du **MIT** présentaient début 2016 leurs travaux sur le chipset Eyeriss utilisant des neurones spécialisés réparties dans 168 cœurs dotés de leur propre mémoire. Mais visiblement sans memristors. L'application visée est la reconnaissance d'images. Le projet est financé par la DARPA.
- Le projet **Nanolitz** aussi financé par la DARPA dans le cadre des projets Atoms to Product (A2P) s'appuie sur des fils microscopiques pour connecter plus efficacement des cœurs et neurones dans des circuits spécialisés.
- La startup californienne **Knowm** (2002, USA) propose depuis 2016 un composant commercial à base de memristors, fabriqué en partenariat avec la Boise State University, à base d'argent ou de cuivre et au prix de $220. Il est destiné en premier lieu aux laboratoires de recherche en réseaux neuronaux.
- Une équipe de chercheurs associant le **CNRS** et **Thales** située à Palaiseau et pilotée par Julie Grollier travaille aussi sur une technologie avancée de processeurs neuromorphiques, en collaboration avec des laboratoires de recherche japonais (AIST Tsukuba) et américains[272].

  Ils utilisent des oscillateurs de résonance magnétique qui permettent de se rapprocher encore plus du mode de fonctionnement des neurones biologiques en facilitant la propagation temporelle des valeurs entre les neurones d'un système, avec des fréquences allant de 300 MHz à 10 GHz, exploitable dans les réseaux de neurones récurrents qui font de la reconnaissance de la parole. Cela permettrait au passage de créer des nano neurones faisant quelques dizaines de nanomètres de large.
- L'**ANR** française a financé le projet collaboratif **MHANN** associant l'INRIA, l'IMS de Bordeaux et Thalès pour créer des memristors ferromagnétiques. Le projet devait être terminé en 2013 et avait bénéficié d'une enveloppe de 740 K€.
- **Rain Neuromorphics** (2018, USA, $2M) est une startup qui met au point une architecture de chipset "Memristive Nanowire Neural Network" [273] qui serait rapide, puissante et très scalable. Les neurones sont connectés entre eux par des nanofils (nanowires) disposés un peu aléatoirement pour relier les neurones entre eux.



  Et cela utilise du deep learning avec la technique du "reservoir network". C'est une approche originale qui mérite le détour.

Ce petit tour des memristors montre que l'IA a du mou sous le pied si le matériel suit. Il suivra à son rythme, mais une chose est sûre : le rapprochement de la mémoire des zones de traitement dans les réseaux de neurones est un point de passage obligé pour décupler leur performance.

---

[271] Voir Molybdenum disulfide memristors; neural network chip for mobile; nanoscale materials for the IoT, de Jesse Allen, février 2016.

[272] Voir Neuromorphic computing with nanoscale spintronic oscillators, janvier 2017 (13 pages).

[273] Ils sont documentés dans la présentation Memristive Nanowire Neural Networks (17 slides) par Jack Kendall, CTO de la startup, et par Juan Nino qui enseigne la physique des matériaux à l'Université de Floride.

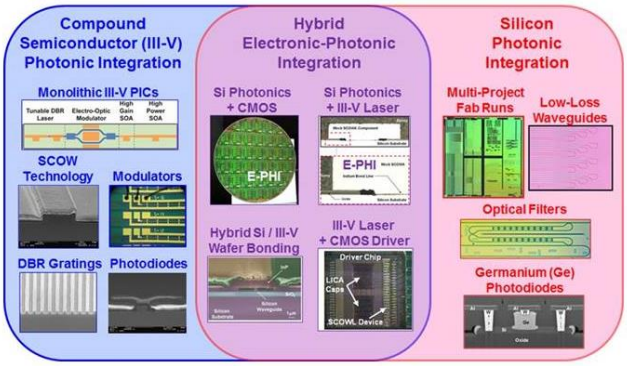
Photonique

La photonique utilise la lumière et les photons pour gérer ou transmettre l'information. Aujourd'hui, dans l'informatique, la photonique est principalement utilisée dans la transmission d'informations sur fibres optiques. Des systèmes de multiplexage permettent de transmettre d'énormes volumes d'information sur une simple fibre optique, jusqu'à des teraoctets par secondes. Elle exploite des composants à base des matériaux dits "III-V"[274] qui associent des éléments correspondent à deux colonnes du tableau de Mendeleev. Aujourd'hui, la photonique est surtout utilisée dans le multiplexage de données sur les liaisons ultra-haut-débit des opérateurs télécoms, dans des applications très spécifiques, ainsi que sur des bus de données optiques de supercalculateurs.

L'intérêt de la photonique est de permettre d'utiliser des fréquences d'horloge énormes, allant jusqu'à 50 GHz. L'un des enjeux se situe dans l'intégration de composants hybrides, ajoutant des briques en photonique au-dessus de composants CMOS plus lents. Intel et quelques autres sont sur le pont.

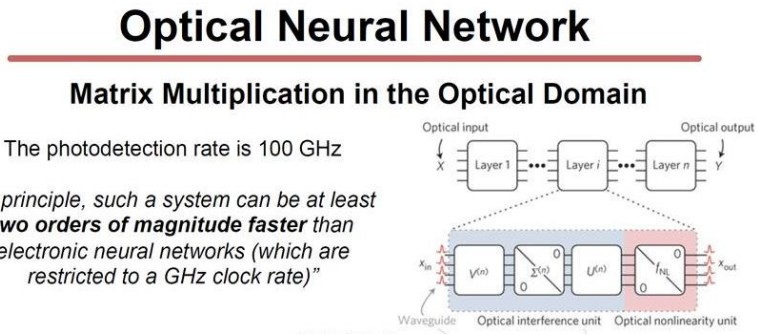
Une fois que l'on aura des processeurs optiques généralistes, il faudra relancer le processus d'intégration. Il est actuellement situé aux alentours de 200 nm pour la photonique et la course se déclenchera alors pour descendre vers 10 à 5 nm comme pour le CMOS actuel.

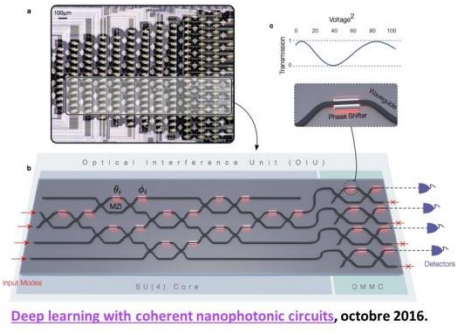
Il existe bien des tentatives de créer des réseaux de neurones en photonique, mais le remplacement de transistors en silicium et technologie CMOS par des transistors en III-V gérant des photons transitant via des fibres optiques est loin d'être facile. Les mécanismes de circulation de la lumière ne sont pas les mêmes que ceux des électrons !



de 4 GHz à 50 GHz

Les premiers chipsets expérimentaux photoniques à réseau de neurones ont fait leur apparition en laboratoire, avec à la clé un potentiel de multiplication de la performance par 50[275] ! **Lightmatter** (2017, USA, $11M) et **Lightelligence** (2017, USA, $10M) sont des spin-offs du projet du MIT qui en est à l'origine et qui testent ces chipsets optiques pour faire de la reconnaissance de la parole, mais seulement pour l'exécution et pas pour l'entraînement[276].





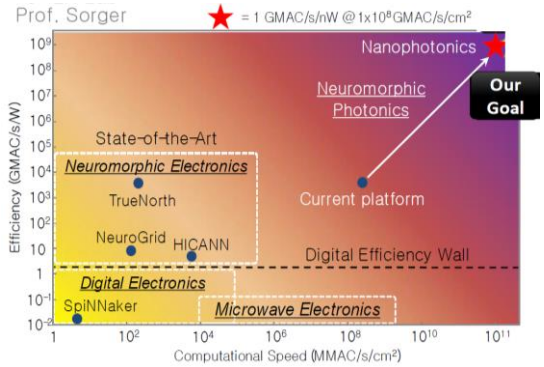
Deep learning with coherent nanophotonic circuits, octobre 2016.

[274] Un sujet que j'avais exploré dans **Comment Alcatel-Lucent augmente les débits d'Internet** en 2013.

[275] Voir Deep learning with coherent nanophotonic circuits de Yichen Shen et al, 2017 (7 pages), cette autre approche de traitement et de stockage optique de l'information réalisée par un laboratoire australien : Storing lightning inside thunder: Researchers are turning optical data into readable soundwaves, septembre 2017 et enfin Photonic Neuromorphic Computing de Rafatul Faria (35 slides).

[276] Voir Light-Powered Computers Brighten AI's Future, juin 2017.

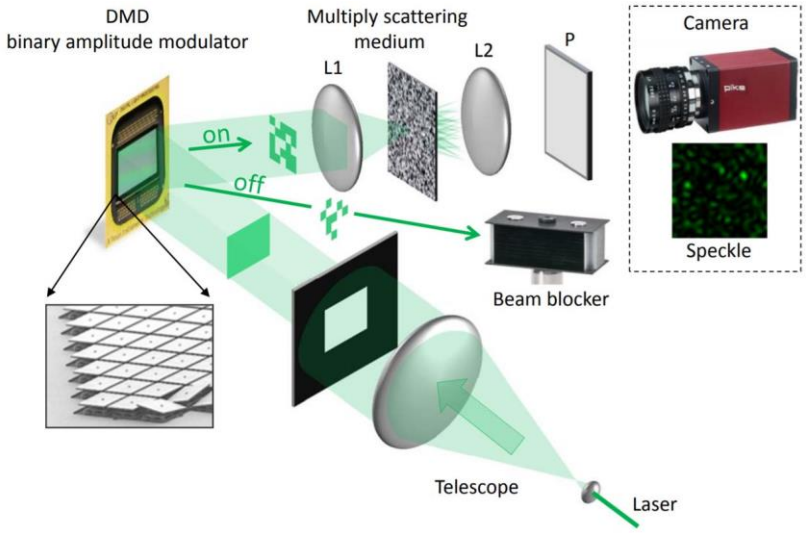
L'équipe de Volker Sorger de **GWU** (Georges Wahington University) met aussi au point des réseaux de neurones en optronique utilisant des composants réalisant des transformées de Fourier rapides[277]. Elle affiche des gains de performance théoriques vertigineux de 10^7 en consommation d'énergie par rapport aux processeurs TrueNorth d'IBM et de 10^9 en rapidité par rapport aux GPU Nvidia !



Deep Learning

Reference	Efficiency (J/MAC)	Speed (MAC/s)
NVIDIA GPU [15]	3.4×10^{-7}	1.7×10^{7}
AlexNet FPGA [16]	2.6×10^{-10}	6.2×10^{10}
UPSIDE Crossbar [17]	1.3×10^{-10}	4.0×10^{10}
UTK Analog Engine [18]	1.0×10^{-12}	11.8×10^{6}
IBM TrueNorth [19]	2.6×10^{-11}	1.3×10^{6}
Nanophotonic Neuromorphic	**7.4×10^{-18}**	**2.0×10^{17}**

Lighton.io (2016, France) met au point un coprocesseur optique servant à accélérer l'entraînement de réseau de neurones sur de gros volumes de données, comme des réseaux convolutionnels. Le procédé est entièrement optique. Un laser émet une lumière qui est agrandie pour éclairer un composant DLP à micro-miroirs comme dans les projecteurs vidéo et un capteur CMOS qui génère des masques.



La lumière passe ensuite au travers d'un filtre de dispersion puis atterri dans un capteur CMOS monochome. Le système s'appuie sur la génération de jeux de données aléatoires permettant de tester simultanément plusieurs hypothèses de calcul, à des fins d'optimisation[278].

Celui-ci récupère le résultat des interférences générées par l'ensemble et un traitement mathématique permet d'en interpréter le résultat. Le dispositif a été miniaturisé progressivement pendant la mise au point, tenant dans l'équivalent d'un serveur 4U à ce stade. La puissance du système vient en particulier de la résolution du DLP et du capteur CMOS, qui est de plusieurs millions de pixels.

Le tout est piloté à partir de bibliothèques Python développées avec TensorFlow. Les applications visées sont en premier lieu la génomique et l'Internet des objets. La startup devrait livrer ses premiers « kits de développement » à des beta testeurs incessamment sous peu.

LightOn n'est pas le seul sur ce créneau avec quelques autres sociétés plus ou moins concurrentes. Mais comme leur communication est bien moins précise que celle de LightOn, il est très difficile de se faire une opinion.

Nous avons donc :

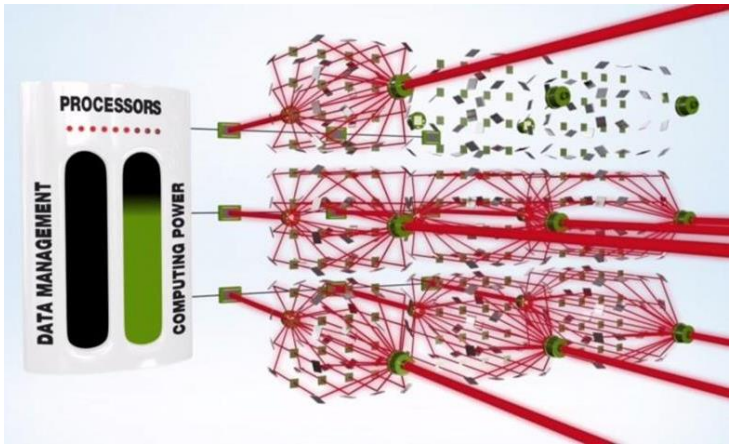
---

[277] Voir aJ/bit Modulators and Photonic Neuromorphic Computing de Volker Sorger, 2017 (15 slides).

[278] Le procédé est décrit dans Random Projections through multiple optical scattering: Approximating kernels at the speed of light, 2015 (6 pages).

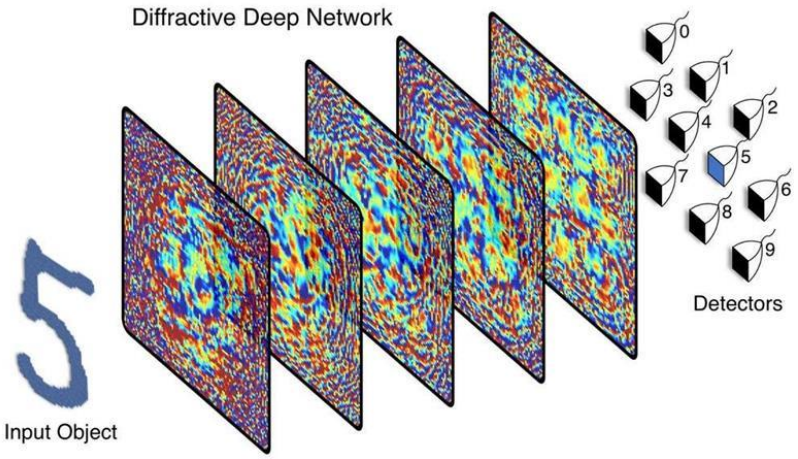
- **Fathom Computing** (2014, USA) qui utilise https://www.fathomcomputing.com/une architecture "électro-optique" qui est capable d'entraîner des réseaux de neurones à mémoire (LSTM) et convolutionnels. Leur Light Processing Unit (LPU) serait capable de lire 90% des tests de la base d'écriture manuscrite MNIST. Dans l'engin, nous avons un laser et des lentilles sans plus de précisions. Le système est adapté à l'algèbre linéaire (ce qui est bizarre lorsque l'on sait que les fonctions d'activation des neurones ne sont généralement pas linéaires) et à la multiplication de matrices qui n'est pas sans rappeller le fonctionnement des ordinateurs quantiques (mais ces derniers utilisent des matrices de nombres complexes). Il leur reste à miniaturiser leur dispositif, ce qui selon eux devrait prendre au moins deux ans. Au passage, ne pas confondre Fathom Computing avec les chipsets Fathom de Movidius, une startup de composant acquise par Intel en 2016 et que nous avons déjà citée précédemment.

- **Optalys** (2013, UK, $5,2M)[279] a de son côté mis en oeuvre son premier réseau convolutionnel début 2018 sur une base MNIST avec 60 000 lettres pour l'entraînement et 10 000 pour les tests. Mais avec un taux de réussite de seulement 70%. Leur système permet aussi de réaliser des transformées de Fourier rapides. Ils sont aussi impliqués dans divers projets, l'un en génétique pour faire des recherches de séquences de génomes.



L'autre pour faire des prévisions météo et un troisième pour la simulation de plasmas et de dynalique des fluides (pour la DARPA).

- Enfin, un projet de l'Université de Los Angeles (**UCLA**) a l'air d'utiliser un procédé optique qui a un lien de parenté avec celui de Lighton[280], s'appuyant sur un système à diffraction multiple réalisé par impression 3D, créant un réseau de neurones de cinq couches de convolution.



Ordinateurs quantiques

J'ai passé tout l'été 2018 à publier une série d'articles sur le sujet, compilés ensuite dans un ebook publié fin septembre[281]. Nous allons ici en extraire ce qui concerne l'intelligence artificielle.

Les ordinateurs quantiques servent à résoudre des problèmes mathématiques dits exponentiels, dont la complexité et la taille grandit exponentiellement en fonction de leur taille. Ils s'appuient sur des qubits, des unités de gestion d'information manipulant des 0 et des 1, mais en état de superposition, et arrangés dans des registres de plusieurs qubits. Un système à base de n qubits est capable de re-

[279] Voir Optalysys - Revolutionary Optical Processing for HPC, septembre 2017 (23 mn).

[280] Voir This 3D-printed AI construct analyzes by bending light, juillet 2018.

[281] Voir L'ebook pour comprendre l'informatique quantique, septembre 2017 (332 pages).

présenter simultanément 2^n états sur lesquels diverses opérations peuvent être appliquées simultanément.

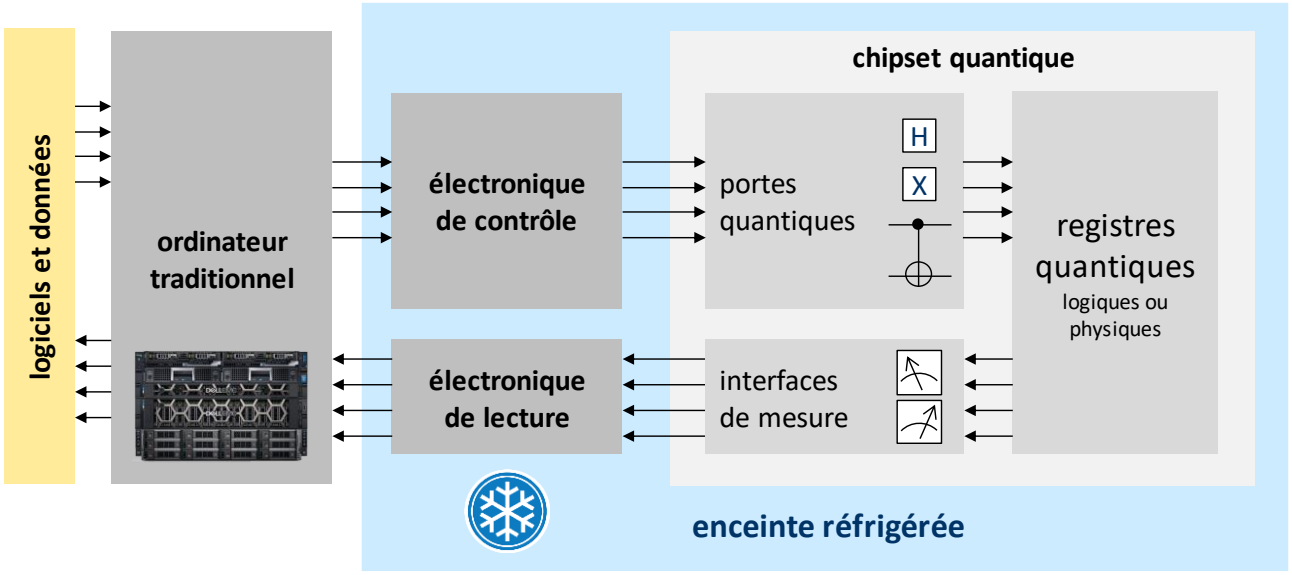
Cependant, les qubits sont binaires lors de leur initialisation (en fait, à 0) et au moment où on lit leur valeur après les calculs. Contrairement à certaines simplifications, les qubits n'ont pas une « capacité de stockage » exponentielle.

Dans de nombreux algorithmes quantiques, au de l'augmentation de la complexité du problème, le temps de calcul augmente de manière linéaire ou polynomiale au lieu de grandir exponentiellement comme dans les ordinateurs traditionnels.

Mais les qubits sont complexes à mettre en œuvre : quelle que soit la technologie employée, ils doivent être souvent refroidis à quelques millikelvins au-dessus du zéro absolu avec des systèmes de cryogénie à base d'hélium liquide. Il est surtout difficile de modifier leur état et de le lire. Les qubits sont sujets aux perturbations de l'environnement ambiant qui génère des erreurs.

Les efforts des physiciens d'aujourd'hui visent à réduire les taux d'erreurs ou à les compenser par des codes de correction d'erreurs. Ceux-ci nécessitent de mettre en œuvre des qubits logiques qui exploitent plusieurs qubits physiques, démultipliant de plusieurs ordres de grandeur le nombre de qubits physiques nécessaires aux calculs.

Un ordinateur quantique est toujours un coprocesseur d'un ordinateur classique, comme décrit dans le schéma suivant[282]. Des portes quantiques physiques agissent sur des qubits qui ne bougent pas physiquement. On enchaine une série de portes quantiques contrôlées par programmation classique sur ces qubits et on lit ensuite le résultat des traitements en évaluant l'état des qubits qui retour un 0 ou un 1.



L'un des premiers algorithmes quantiques apparus qui soit traitable par un ordinateur quantique est celui de **Peter Shor** (AT&T), en 1994. Il permet de factoriser des nombres entiers en nombres premiers avec un temps de calcul qui évolue en fonction du logarithme du nombre plutôt de sa racine carrée dans les calculateurs traditionnels[283]. Il permet de casser les clés publiques utilisées en cryptographie, en particulier avec l'algorithme RSA qui est omniprésent sur Internet. Ce qui le remet

[282] J'en détaille toutes les composantes dans Comprendre l'informatique quantique - ordinateur quantique, juillet 2018

[283] La version 2017 de cet ouvrage contenait une grosse erreur à ce sujet. J'écrivais au sujet de l'algorithme de Shor « Il permet de factoriser des nombres entiers en nombres premiers avec un temps de calcul qui évolue en fonction du logarithme du nombre plutôt de son exponentielle comme avec les calculateurs traditionnels ». Ce n'était pas « exponentielle » mais « racine carrée ». Nuance !

sérieusement en question[284] ! Ont suivi divers algorithmes de recherche (1996), d'optimisation (parcours du voyageur de commerce), de simulation de la physique des matériaux et même des mécanismes de la photosynthèse.

Une classe entière d'algorithmes quantiques est dédiée à l'optimisation dans les domaines du machine learning et du deep learning.

L'un de ces algorithmes relève de l'entraînement de réseaux de neurones. Mais cela ne change pas les fonctionnalités accessibles, qui sont déjà exploitables. Cela ne jouera un rôle que le jour où on alignera des millions de qubits avec un faible bruit.

Cela va d'ailleurs poser des problèmes d'explicabilité des algorithmes car on ne pourra pas expliquer du tout le résultat, car on ne pourra pas forcément lire le poids des synapses ou décomposer le processus de ces réseaux de neurones quantiques.

Ces algorithmes de réseaux de neurones doivent contourner le fait que les fonctions d'activation des neurones sont généralement non linéaires, comme les sigmoïdes qui sont couramment utilisées alors que les portes quantiques appliquent toutes des transformations linéaires[285].

Ces techniques seront concurrencées par les futurs processeurs neuromorphiques à base de memristors qui permettront de faire converger plus rapidement les réseaux par rétropropagation, comme vu précédemment.

Dans Quantum Machine Learning, mai 2018 (24 pages), on trouve ce tableau qui positionne clairement les différentes accélérations quantiques associées à divers algorithmes utilisés dans le machine learning et le deep learning. Les accélérations en log(N) sont plus importantes que celles qui sont exprimées en racine carré de N.

Ce document évoque de nombreux algorithmes quantiques de bas niveau qui sont très utiles au machine learning et au deep learning : le qBLAS (quantum basic linear algebra subroutines), la résolution d'équations linéaires, la descente de gradient pour la rétropropagation dans l'entraînement des réseaux de neurones, la PCA pour déterminer les variables clés d'un jeu de données (Principal Component Analysis, utilisée dans le machine learning traditionnel) et le SVM (support vector machine, une méthode traditionnelle de segmentation dans le machine learning). Le tout, avec une amélioration exponentielle de vitesse de traitement.

Method	Speedup	AA	HHL	Adiabatic	QRAM
Bayesian Inference [107, 108]	$O(\sqrt{N})$	Y	Y	N	N
Online Perceptron [109]	$O(\sqrt{N})$	Y	N	N	optional
Least squares fitting [9]	$O(\log N^{(*)})$	Y	Y	N	Y
Classical BM [20]	$O(\sqrt{N})$	Y/N	optional/N	N/Y	optional
Quantum BM [22, 62]	$O(\log N^{(*)})$	optional/N	N	N/Y	N
Quantum PCA [11]	$O(\log N^{(*)})$	N	Y	N	optional
Quantum SVM [13]	$O(\log N^{(*)})$	N	Y	N	Y
Quantum reinforcement learning [30]	$O(\sqrt{N})$	Y	N	N	N

On retrouve cette liste d'algorithmes de machine learning en version quantique dans Quantum Machine Learning What Quantum Computing Means to Data Mining de Peter Wittek, 2014 (178 pages)[286].

---

[284] Voir Comprendre l'informatique quantique – cryptographie, août 2018.

[285] L'astuce est expliquée dans Quantum Neuron: an elementary building block for machine learning on quantum computers, de Yudong Cao, Gian Giacomo Guerreschi et Alan Aspuru-Guzik en 2017 (30 pages).

[286] Côté sources d'information sur ce sujet, voir notamment Application of Quantum Annealing to Training of Deep Neural Networks. (2015), On the Challenges of Physical Implementations of RBMs, 2014, avec notamment Yoshua Bengio et Ian Goodfellow parmi les auteurs, illustrant l'intérêt des spécialistes de l'IA pour le quantique et Quantum Deep Learning, 2014, le tout étant extrait de Near-Term Applications of Quantum Annealing, 2016, Lockheed Martin (34 slides).

Il existe même des algorithmes quantiques de GAN (Generative Adversarial Networks), pour ces réseaux de neurones qui génèrent des contenus synthétiques à partir de contenus existants en vérifiant leur plausibilité via un réseau de neurones de reconnaissance[287].

Table 1.1 The Characteristics of the Main Approaches to Quantum Machine Learning

Algorithm	Reference	Grover	Speedup	Quantum Data	Generalization Performance	Implementation
K-medians	Aïmeur et al. (2013)	Yes	Quadratic	No	No	No
Hierarchical clustering	Aïmeur et al. (2013)	Yes	Quadratic	No	No	No
K-means	Lloyd et al. (2013a)	Optional	Exponential	Yes	No	No
Principal components	Lloyd et al. (2013b)	No	Exponential	Yes	No	No
Associative memory	Ventura and Martinez (2000)	Yes		No	No	No
	Trugenberger (2001)	No		No	No	No
Neural networks	Narayanan and Menneer (2000)	Yes		No	Numerical	Yes
Support vector machines	Anguita et al. (2003)	Yes	Quadratic	No	Analytical	No
	Rebentrost et al. (2013)	No	Exponential	Yes	No	No
Nearest neighbors	Wiebe et al. (2014)	Yes	Quadratic	No	Numerical	No
Regression	Bisio et al. (2010)	No		Yes	No	No
Boosting	Neven et al. (2009)	No	Quadratic	No	Analytical	Yes

Il existe de nombreuses catégories de processeurs quantiques qui se définissent par leur technologie de qubits.

Les principales sont notamment à base de :

- **Recuit simulé quantique**, ou quantum annealing, chez le canadien **D-Wave** (1999 $174M) qui est le seul à commercialiser des ordinateurs quantiques à ce jour, avec 2048 qubits, même si leur efficacité est contestée. Pour les puristes, ce ne sont pas véritablement des ordinateurs quantiques. Mais ils ont l'avantage d'exister. Et de nombreux algorithmes quantiques de machine learning fonctionnent dessus.
- **Boucles supraconductrices**, ou superconducting loops, chez **IBM**, **Google, Rigetti** ainsi qu'au **CEA** de Saclay en France. Le record en date est situé autour de 50 qubits.
- **Qubits topologiques**, chez **Microsoft** avec les fermions de Majorana dont l'existence a été plus ou moins prouvée en laboratoire, et dans les Bell Labs de **Nokia** aux USA. Aucune démonstration d'ordinateur quantique de ce type, même avec un seul qubit n'a pour l'instant été réalisée.
- **Quantum dots sur silicium**, chez **Intel**, ainsi qu'au **CEA-Leti** à Grenoble. Ils présentent l'avantage d'être miniaturisables et donc, de « scaler », ce qui n'est pas le cas de la plupart des autres technologies de qubits. Mais ils sont toujours en cours de mise au point.
- **Ions piégés**, comme chez la startup **ionQ** (2016, USA, $20M) issue de l'Université du Maryland et de l'Université Duke en Caroline du Nord. Elle aligne aussi une cinquantaine de qubits.

Certains types de qubits sont notamment plus difficiles à stabiliser que d'autres. Les caractéristiques qui déterminent la performance d'un ordinateur quantique sont nombreuses : la première est le taux d'erreur des qubits et des portes qui agissent sur eux, la seconde est le temps de cohérence des qubits, c'est-à-dire, le temps pendant lequel les qubits sont en état de superposition, avec le temps d'exécution des portes quantiques, ces temps conditionnent le nombre de portes quantiques qui pourront être enchainées dans un algorithme.

Pour faire du machine learning quantique à grande échelle, il faudrait disposer de milliers de qubits stables, et on en est encore loin. Lorsque les ordinateurs quantiques verront le jour et, surtout, seront réellement programmables et généralistes, il est probable que l'on assistera à une explosion de leurs domaines d'applications. A quelle échéance ? L'incertitude est très grande à ce sujet avec des prévi-

[287] C'est documenté dans Quantum generative adversarial learning de Seth Lloyd et Christian Weedbrook, 2018 (5 pages).

sions qui s'étalent entre 5 ans et… jamais ! L'échéance de la singularité par l'AGI est plus déterministe de ce point de vue-là, même si les prévisions associées (2045 ou autre) ne veulent rien dire.

Mémoire

Les GPU et les processeurs neuromorphiques sont d'autant plus performants dans les phases d'entraînement qu'ils accèdent rapidement aux données en mémoire, et notamment aux paramètres des réseaux de neurones qui peuvent être des dizaines de millions voire des milliards de variables à ajuster très fréquemment.

Pour cela, les technologies matérielles rapprochent de plus en plus les unités de traitement de mémoires de plus en plus rapides. Un serveur peut avoir jusqu'à une demi-douzaine de niveaux de mémoire qui optimiseront la performance de l'ensemble. Sachant que plus la mémoire est rapide, plus elle est couteuse et plus sa taille est limitée. Nous avons donc une hiérachie de mémoires dont la vitesse augmente inversement proportionnellement à leur taille.

Le cas le plus extrême est celui d'une mémoire non volatile directement intégrée au sein des neurones comme nous l'avons creusé dans la partie sur les memristors. Pour certains scientifiques, c'est la seule voie de salut pour continuer à faire progresser les architectures matérielles de l'IA[288].

Nous allons faire le tour des principaux types et niveaux de mémoire qui équipent aujourd'hui les serveurs. Nous nous focalisons sur les serveurs qui réalisent l'entraînement de réseaux de neurones car ils sont les plus sollicités en calcul. L'inférence d'un réseau de neurones déjà entraîné est très simple et peu consommateur de calcul par rapport à son entraînement.

Mémoire cache

Au sein des processeurs se trouve de la mémoire cache volatile qui est utilisée directement par les unités de traitement. En technologie CMOS classique, sa vitesse d'accès est ce qui se fait de plus rapide, et dépasse les To/s (téra-octets par secondes).

Un processeur courant comprend précisément deux à trois niveaux de cache et des registres mémoires. Plus on se rapproche des unités de traitement, plus l'accès à cette mémoire cache est rapide, mais plus elle est limitée en capacité, de l'ordre de quelques dizaines de Ko, soit juste de quoi alimenter les registres de calculs utilisés dans les processeurs et de quoi en lire les résultats.

La mémoire intégrée dans les processeurs est limitée à quelques Mo. Pourquoi donc ? Parce qu'elle est chère à fabriquer et qu'il n'y a pas beaucoup de place sur les puces des processeurs.

HBM2 et HMC

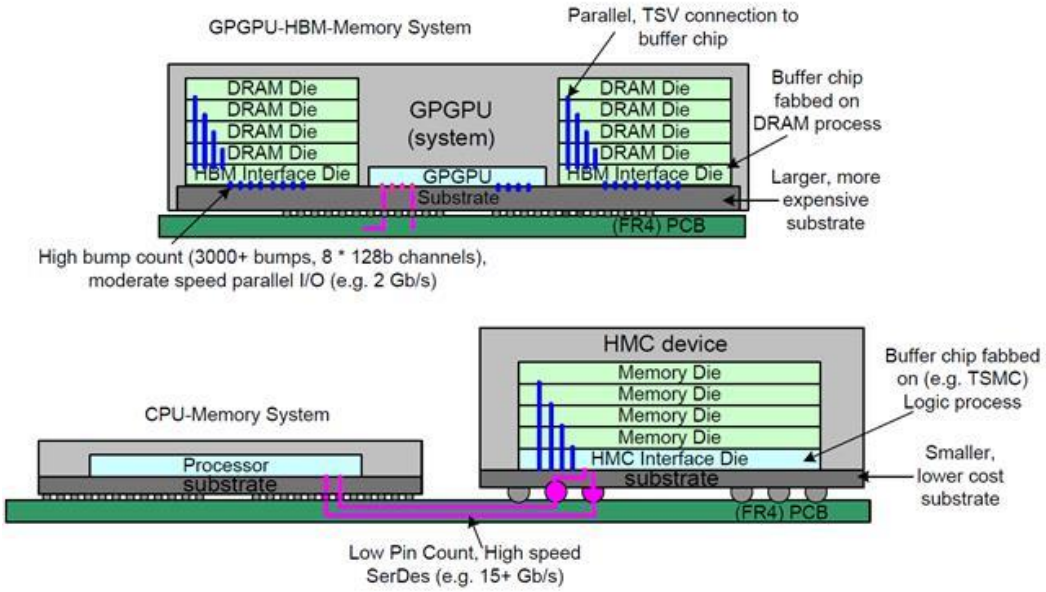
Autour des GPU et de certains TPU (Tensor Processing Units) se trouve maintenant souvent une mémoire complémentaire très rapide utilisant l'un des deux grands standards du marché **HBM2** (High Bandwidth Memory) ou **HMC** (Hyper Memory Cube).

Le premier standard est promu par AMD et le Coréen SK Hynix et le second par l'Américain Micron avec le support de Samsung. Cette mémoire qui atteint aujourd'hui 16 Go à 32 Go est située dans des circuits intégrés empilés pas paquets de 4 ou 8 et reliés entre eux et avec le GPU ou le CPU par des micro-conducteurs métalliques. Cette proximité permet d'avoir une vitesse rapide de transfert d'information entre la mémoire et le processeur.

L'intégration avec le GPU est plus étroite pour le HBM2 car la mémoire et le GPU sont installés sur un substrat commun tandis que pour le HMC, la mémoire est placée sur la carte mère au même titre

[288] Voir AI Architectures Must Change de Brian Bailey, août 2018 et 3D Neuromorphic Architectures, de Katherine Derbyshire, décembre 2017.

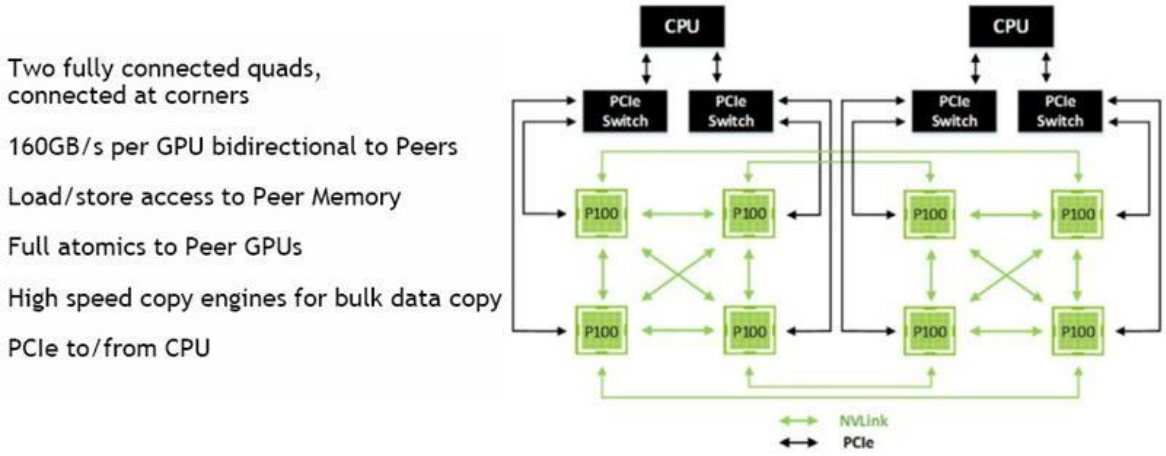
que le CPU[289]. Ces mémoires permettent d'atteindre des débits allant jusqu'à 900 Go/s dans le Nvidia Volta V100 lancé en 2017.



HBM2 est utilisé dans les GPU Nvidia V100 et HMC l'est dans les processeurs serveurs Intel Xeon Phi ainsi que dans les FPGA Intel Stratix 10MX utilisés notamment par Microsoft dans ses processeurs neuromorphiques Brainwave.

NVLink

La technologie **NVLink** de Nvidia permet de relier les GPU entre eux ou les GPU avec les CPU à une vitesse atteignant 300 Go/s par composant. La version 2.0 de NVLink utilisé des chipsets dédiés, les NVSwitch, donnant un déébit total de 900 Go/s.



Cette connexion permet de répartir optimalement les traitements parallèles sur plusieurs processeurs. En effet, les grands modèles de réseaux de neurones doivent être répartis sur plusieurs GPU et plusieurs serveurs. Ils peuvent être des milliers !

[289] Source du schéma qui suit : A Talk on Memory Buffers, Inphi.

Infiniband

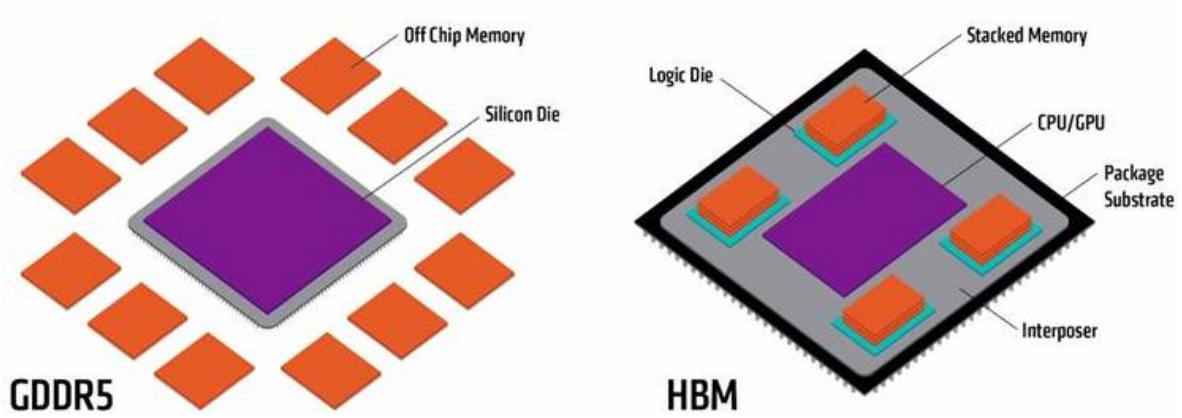
Infiniband est une technologie permettant de relier les serveurs entre eux avec des débits compris entre 100 et 200 Gbits/s. La connexion se fait via un câble différent du RJ45 des réseaux Ethernet.

Les composants supportant Infiniband sont commercialisés par **Mellanox Technologies** (1999, Israël, $89M) et par Intel. Infiniband est notamment complété par le standard RoCE (RDMA over Converged Ethernet) qui permet de permettre l'accès à la mémoire d'un serveur par un autre serveur. Infiniband est concurrencé par Fibre Channel, une autre technologie de liaison entre serveurs qui peut atteindre 128 Gbits/s et sert surtout à l'optimisation de l'accès au stockage. Toutes ces technologies sont utilisées dans les data centers et les super-ordinateurs (HPC).

GDDR

La mémoire **GDDR** est utilisée dans les cartes graphiques et est plus rapide que la mémoire DDR4 qui est utilisée actuellement dans les micro-ordinateurs. Elle atteint une bande passante de 48 Go/s dans sa version 5.0. Les chipsets en version 6.0, plus rapides, sont apparus en 2018.

Nvidia a intégré de la mémoire GDDR 6 dans ses cartes graphiques GeForce RTX 2080 Ti, RTX 2080 et RTX 2070 en architecture « Turing », lancées en août 2018. Celles-ci visent surtout le marché du jeu vidéo et se distingent par leurs capacité de « ray tracing » temps-réel pour la génération d'images 3D photoréalistes, mais comprennent tout de même des tenseurs pour l'exécution d'applications de deep learning. Le chipset peut exécuter 500 trillions d'opérations par seconde dans ces tensors, un niveau difficile à comparer avec d'autres processeurs[290].



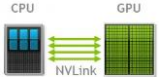
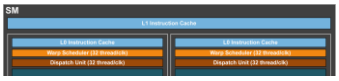
La mémoire GDDR est remplacée par de la mémoire HMC ou HBM2 depuis quelques temps dans les GPU haut de gamme.

Ces nombreuses avancées montrent que les fabricants de ces composants ont encore du mou sous la pédale. Peu d'obstacles les empêchent ainsi, à moyen terme, d'intégrer de plus grandes capacités de mémoire rapide dans les processeurs eux-mêmes. C'est une question de maîtrise de la fabrication de circuits intégrés de grande taille et avec des dizaines de milliards de transistors.

Citons pour terminer sur la mémoire un cas original avec la startup fabless **Upmem** (2015, France, 3M€), basée à Grenoble, la capitale française des nanotechnologies, qui intègre des unités de traitement DPU (DRAM Processing Units) directement à l'intérieur de mémoires DRAM, permettant

[290] Voir les détails sur les tenseurs des GPU Turing : The NVIDIA Turing GPU Architecture Deep Dive: Prelude to GeForce RTX, septembre 2018. Ils supportent 114 Teraflops en FP16 (flottant 16 bits) et 455 teraops en entiers 4 bits. Le processeur comprend 18,9 milliards de transistors, un peu moins que les 21 milliards du V100 lancé en 2017 et destiné aux serveurs.

une accélération de certains traitements applicables notamment au data mining. L'idée, issue du CEA-LETI, consiste à intégrer dans des chipsets de mémoire des unités de traitement RISC (jeu d'instruction simple) 32 bits, le dimensionnement pouvant être de 256 unités de traitement dans des chipsets de 16 Go de RAM. Bref, au lieu de mettre de la mémoire rapide dans des processeurs, ils mettent des unités de traitement dans la mémoire rapide ! Ces DRAM actives sont des coprocesseurs de traitement de CPU traditionnels. Reste à les fabriquer en volume et à les faire intégrer dans des serveurs par leurs constructeurs !

		vitesse max	capacité
SSD M.2 PCIe stockage		3 Go/s	>1 To
DDR4 mémoire externe CPU		3,2 Go/s	>16 Go
Infiniband comm interserveur		25 Gos/s	*bus de données*
GDDR6 mémoire externe GPU		672 Go/s	2 Go – 12 Go
NVLink 2.0 comm inter-GPU/CPU		900 Go/s	*bus de données*
HBM2 / HMC mémoire externe GPU		900 Go/s	16 Go
GPU cache & registres mémoire interne GPU		> 16 To / s	6 Mo (L1)

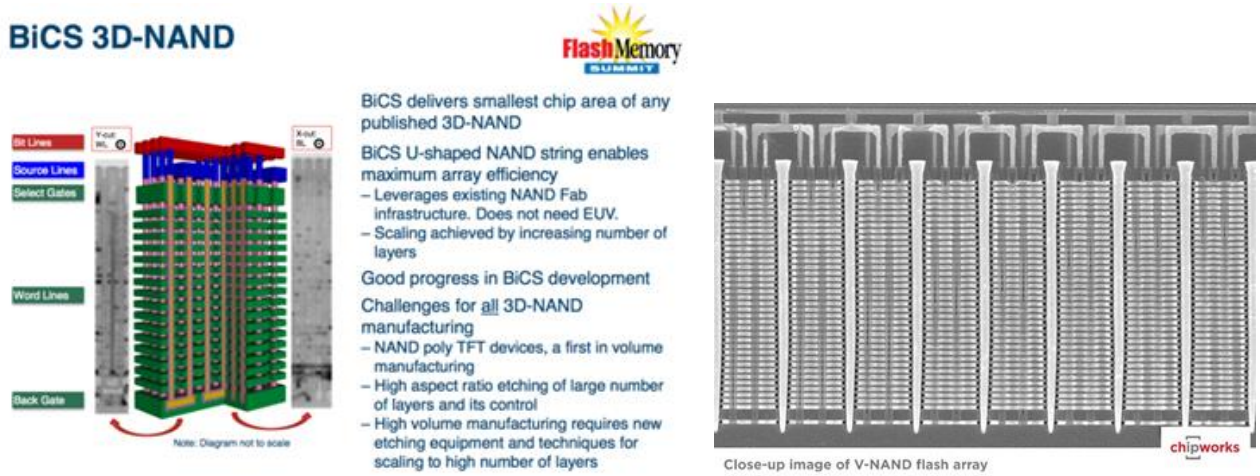
Stockage

Le stockage d'information se fait de plus en plus dans des SSD, les disques de stockage sans plateau mobile et à base de mémoire flash NAND et V-NAND. Ce sont des circuits intégrés à plusieurs couches, empilant jusqu'à 72 couches de transistors. Les SSD grand public atteignent aujourd'hui une capacité de 2 To avec une vitesse d'accès de 3 Go/s dans le meilleur des cas (sur certains laptops Apple). Il existe des SSD de bien plus grande capacité qui sont destinés aux serveurs de datacenters.

Si la loi de Moore a tendance à se calmer du côté des processeurs CMOS, elle continue de s'appliquer au stockage. Elle s'est appliquée de manière plutôt stable aux disques durs jusqu'à présent. Le premier disque de 1 To (Hitachi en 3,5 pouces) est apparu en 2009 et on en est maintenant à 14 To. Donc, plus que 2 puissance 4 et la loi de Moore est sauve !

Le progrès s'est ensuite déplacé vers les disques SSD à mémoires NAND dont la capacité augmente régulièrement tout comme sa vitesse d'accès et le tout avec une baisse régulière des prix.

Les perspectives de croissance sont ici plus optimistes qu'avec les processeurs CMOS. Les records côté SSD sont de 30n72 To chez Samsung (en 2,5 pouces, février 2018) et de 100 To chez Nimbus Data (en 3,5 pouces, mars 2018).



L'augmentation de la densité des mémoires NAND profite des architectures en trois dimensions qui sont maintenant courantes, comme avec les V-NAND de **Samsung** qui sont utilisées dans leurs SSD pour laptops, desktops et serveurs. Nous avons aussi **Toshiba** (*ci-dessus*) avec sa technologie BiCS. Les puces de mémoire 3D comprennent avec plusieurs couches empilées de transistors (*ci-dessus* à droite), ou de transistors montés en colonnes.

Le niveau d'intégration le plus bas des transistors est ici équivalent à celui des CPU les plus denses : il descend jusqu'à 10 nm. On sait empiler aujourd'hui jusqu'à 64 couches de transistors, et cela pourrait rapidement atteindre une centaine de couches.

La technologie 3D XPoint d'**Intel** et **Micron** qui combine le stockage longue durée et une vitesse d'accès équivalente à celle la mémoire RAM associée aux processeurs est aussi prometteuse même si elle connaît un double retard à l'allumage : côté disponibilité comme côté performance.

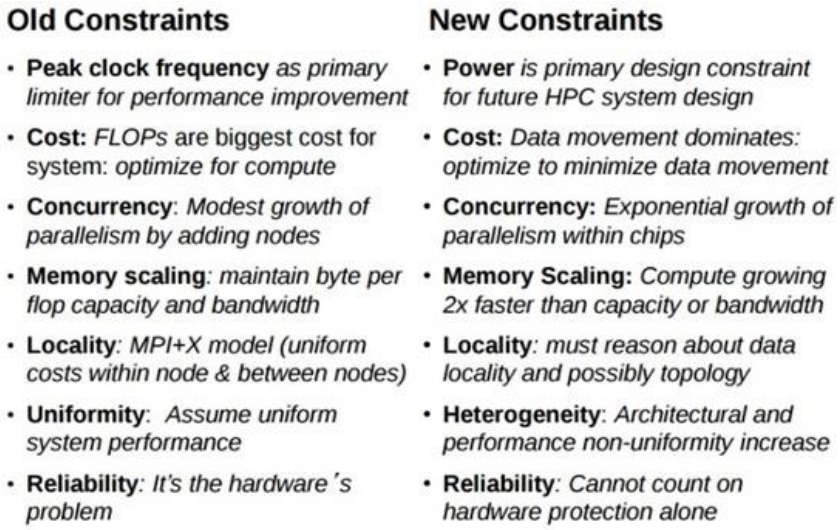
Pourquoi cette intégration verticale est-elle possible pour la mémoire et pas encore pour les processeurs (GPU, CPU) ? C'est lié à la résistance à la montée en température. Dans un processeur, une bonne part des transistors fonctionne en même temps alors que l'accès à la mémoire est séquentiel et donc n'active pas simultanément les transistors. Un processeur chauffe donc plus qu'une mémoire. Si on empilait plusieurs couches de transistors dans un processeur, il se mettrait à chauffer bien trop et s'endommagerait. Par contre, on sait assembler des circuits les uns sur les autres pour répondre aux besoins d'applications spécifiques.

Ce modèle de mémoire en 3D est également appliqué à la RAM, notamment par l'américain **Micron** avec sa technologie Hyper Memory Cube. Pour les supercalculateurs, une tâche ardue est à accomplir : accélérer la vitesse de transfer des données du stockage vers les processeurs au gré de l'augmentation de la performance de ces derniers. Cela va aller jusqu'à intégrer de la connectique à plusieurs centaines de Gbits/s dans les processeurs. Mais la mémoire ne suit pas forcément.

Aujourd'hui, un SSD connecté en PCI et avec un connecteur M.2 est capable de lire les données à la vitesse vertigineuse de 3,2 Go/s, soit un dixième de ce qui est recherché dans les calculateurs à haute performance (HPC). Avec 3D XPoint, l'accès aux données serait 1000 fois plus rapide qu'avec les SSD actuels, modulo l'interface utilisée. La technologie aura probablement un impact important pour les systèmes d'IA temps réel comme IBM Watson. Rappelons-nous que pour Jeopardy, l'ensemble de la base de connaissance était chargée en mémoire RAM pour permettre un traitement rapide des questions [291]!

[291] IBM Watson avait chargé en mémoire tout Wikipedia et les questions lui étaient soumises par écrit et pas oralement. Bref, Watson et les joueurs homo-sapiens ne jouaient vraiment pas à armes égales !

Cette augmentation de la rapidité d'accès à la mémoire, qu'elle soit vive ou de longue durée, est indispensable pour suivre les évolutions à venir de la puissance des processeurs avec l'une des techniques que nous avons examinées juste avant. Et d'autres obstacles sont à surmonter [292] ! Cela rappelle le besoin d'équilibrer les architectures de systèmes de plus en plus performants.



Des chercheurs d'université et même chez Microsoft Research cherchent à **stocker l'information dans de l'ADN**, en partenariat avec la startup américaine Twist Bioscience créé par la française Emily Leproust. Les premières expériences menées depuis quelques années sont prometteuses[293]. La densité d'un tel stockage serait énorme. Son avantage est sa durabilité, estimée à des dizaines de milliers d'années, voire plus selon les techniques de préservation. Reste à trouver le moyen d'écrire et de lire dans de l'ADN à une vitesse raisonnable.

Aujourd'hui, on sait imprimer des bases d'ADN à une vitesse incommensurablement lente par rapport aux besoins des ordinateurs. Cela se chiffre en centaines de bases par heure au grand maximum. Cette vitesse s'accélèrera sans doute dans les années à venir.

Mais, comme c'est de la chimie, elle sera probablement plus lente que les changements de phase ou de magnétisme qui ont court dans les systèmes de stockage numérique actuels. La loi de Moore patientera donc quelques décennies de ce côté là, tout du moins pour ses applications dans le cadre de l'IA.

Big data, data lake et cloud

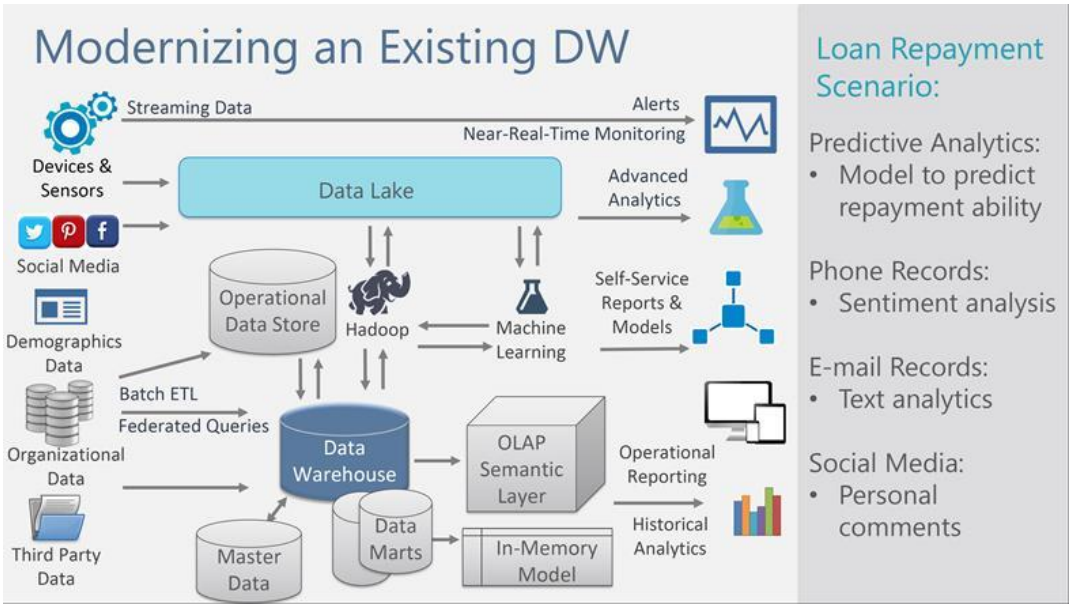
Le big data est étroitement lié à l'intelligence artificielle. C'est même d'une certaine manière le sang qui l'alimente. Nombre d'applications de machine learning et de deep learning exploitent de gros volumes de données internes aux entreprises. Plus l'entreprise détient de points de contacts avec des infrastructures ou des clients, plus volumiques sont les données captées exploitables par des solutions de machine learning.

Les infrastructures de big data sont donc clés pour alimenter les applications de l'IA. Les données consolidées étaient peut-être faiblement exploitées jusqu'à présent et le seront mieux grâce à la puissance des outils du machine learning. Celui-ci est d'ailleurs souvent présenté comme l'aboutissement des outils de *data analytics* et de *data intelligence*.

---

[292] Source du schéma : Why we need Exascale and why we won't get there by 2020, 2014 (56 pages).

[293] Sachant néanmoins qu'elles ont démarré en 1994 avec les travaux de Leonard M. Adleman aux USA, documentés dans Computing with DNA paru dans Scientific American en 1998. A cette époque, Adleman voulait créer un ordinateur à base d'ADN. Mais sa conclusion était que l'ADN était surtout un moyen intéressant de stockage de gros volumes d'information. J'ai remarqué au passage dans l'article que le cout de la génération de molécules d'ADN était déjà relativement bas à cette époque : $1,25 la paire de bases d'ADN. Il démarre en 2016 à $0,2, soit seulement 6 fois moins. En plus de 20 ans ! Encore un exemple où la loi de Moore ne s'est pas du tout appliquée. Pour l'instant !



La consolidation des données se met maintenant en place dans le cadre de la création de **data lakes**[294], des entrepots de données généralistes pour l'entreprise. Cela simplifie leur exploitation par les applications, notamment celles du machine learning qui peuvent exploiter plusieurs sources de données différentes et complémentaires.

Un data lake contient des données brutes tandis qu'un entrepôt de données classique contient des données déjà traitées et agrégées. Les sources des données peuvent être multiples. Les données sont structurées (bases relationnelles SQL), faiblement structurées (CSV, logs, données issues de capteurs et objets connectés) et non structurées (documents, mails, images). Les données peuvent être conservées pour être utilisées plus tard. On ne connaît pas a priori leur utilisation lorsqu'on les accumule. Ce qui suppose tout de même que l'on en a tout de même une petite idée !

Elles sont physiquement stockées dans l'entreprise ou dans le cloud dans des bases de données distribuées (comme avec HDFS, Hadoop Distributed File Storage), dans des bases NoSQL ou dans des bases objets telles que Amazon S3 ou Azure Blog Storage.

Un datalake est construit avec des outils d'ingestion de données, qui extraient données et métadonnées des bases qui l'alimentent, des outils de stockage souvent distribués et des moyens de recherche et d'extraction de données qui interrogent les bases avec des outils relationnels ou « no SQL ». Un data lake « connaissance des clients » va ainsi conserver les données issus de systèmes disparates : CRM, SFA, centres d'appels, support technique et réseaux sociaux.

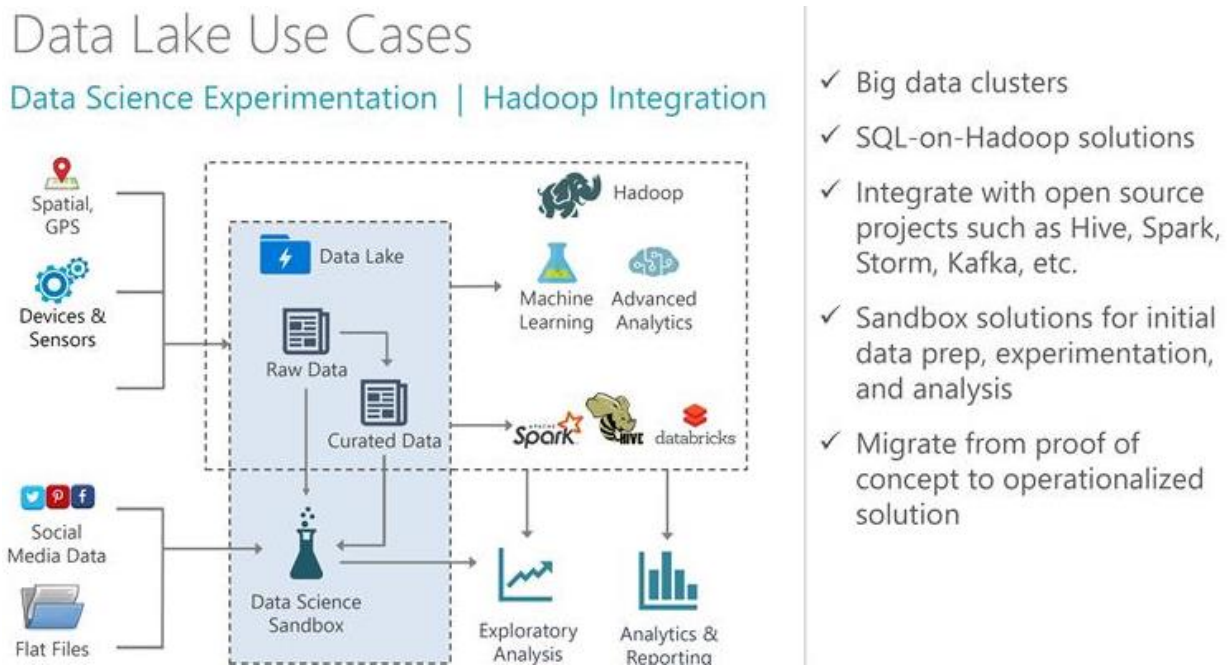
Le tout n'est évidemment accessible que via une couche de sécurité et de droits d'accès. Les data lakes sont utilisés pour l'analyse de données en mode lecture par opposition aux systèmes transactionnels qui fonctionnent en lecture/écriture et en temps réel[295].

Les données extraites du data lake doivent être préparées et netoyées avant d'alimenter l'entraînement de solutions de machine learning.

---

[294] Le concept de data lake a été évoqué pour la première fois par Dorian Pyle en 1999 dans Data preparation and data mining (466 pages) et promu ensuite en 2010 par James Dixon de la société américaine Pentaho. Voir Pentaho, Hadoop, and Data Lakes de James Dixon.

[295] Informations et schéma issus de Designing a Modern Data Warehouse + Data Lake de Melissa Coates, 2017 (72 slides). Voir aussi ce cours de Omar Boussaid de l'Université de Lyon, Du Data Warehouse au… Data Lake, 2017 (54 slides).



Elles peuvent être exploitées par **Spark** de la fondation Apache, qui fonctionne en mode distribué sur des données stockées en RAM dans les serveurs, ce qui permet d'entraîner des modèles de machine learning très rapidement via sa bibliothèque **SparkML**[296]. D'autres bibliothèques supportent le développement d'applications de machine learning distribué, comme **Mlbase**[297], **GraphLab** et **TensorFlow**.

Les infrastructures de big data profitent de divers progrès dans le stockage, le parallélisme et la communication entre serveurs. D'ailleurs, le marché du big data et de l'analytics a rapidement sauté dans le bain de l'IA, même si cela comporte probablement pas mal d'IA washing. L'offre est en tout cas plus qu'abondante.

Certains traitements restent allergiques au calcul distribué traditionnel en cluster. Par exemple dans le cas du deep learning, notamment appliqué à la reconnaissance d'images, il sera fortement recommandé de répartir les traitements localement via des pools de cartes graphiques embarqués sur les mêmes machines pour minimiser les échanges réseau. Ces derniers étant susceptibles de faire exploser les temps d'exécution.

Le champ de la gestion des données entre big data, data lake, cloud et machine learning est couvert par une offre des grands acteurs de l'IT (Oracle, Microsoft, IBM, …) ainsi que de nombreuses startups. En voici quelques-unes de notables :

- **Bigstream** (2015, USA, $5,5M) fournit une plateforme matérielle et logicielle pour gérer l'extraction de donnée, leur intégration et leur analyse. Leur Hyper-acceleration Layer permet de faire tourner les différentes briques logicielles, dont TensorFlow sur les meilleures plateformes matérielles (CPU, GPU, FPGA) pour optimiser leur performance. Ils ciblent les marchés de la publicité et de la finance.
- **Databricks** (2013, USA, $247M) propose sa plateforme Mlflow qui gère l'administration de ses processus de machine learning. C'est une sorte d'outil de gestion du workflow de ses processus de ML, avec l'ingestion de données, les transformations, le data cleaning et l'exploitation d'algorithmes de machine learning. L'offre de Databricks est intégrée dans Microsoft Azure depuis 2017.

---

[296] Voir Data Lakes A Solution or a new Challenge for Big Data Integration de Christoph Quix de l'Institut Fraunhofer, 2016 (28 slides) et Azure Data Lake What Why and How de Melissa Coates, septembre 2018 (81 slides).

[297] Voir MLbase: A Distributed Machine-learning System (7 pages).

- **Saagie** (2013, France, $11,6M) propose une plateforme open source de gestion de ses données dans le cloud utilisables notamment par vos applications de machine learning. C'est du backoffice qui gère les flux de données de l'entreprise, dont son data lake, gérant les accès utilisateurs, les processus d'ingestion de données, le respect du RGPD et la supervision. La startup vise en premier le marché de la finance.

- **Tamr** (2014, USA, $59,2M) est une autre startup qui propose des outils d'ingestion de données disparates pour les applications de big data et de machine learning qui exploite en amont de l'expertise humaine, donc pas scalable, pour le nettoyage des données. La startup exploite des recherches du MIT Computer Science and Information Lab (CSAIL).
- **Sentient Technologies** (2007, USA, $174M) développe une solution d'IA massivement distribuable sur des millions de CPUs, visant les marchés de la santé, de la détection de fraudes et du du e-commerce. La société dit employer des méthodes d'IA avancées pour détecter des tendances dans les données. C'est du "big data" revisité. Le système imite les processus biologiques pour faire de l'auto-apprentissage. On trouve des morceaux de deep learning et des agents intelligents dedans. Ces agents sont évalués avec des jeux de tests et les meilleurs conservés tandis que les plus mauvais sont éliminés. Bref, c'est une sorte de Skynet. L'un des fondateurs de la société est français, Antoine Blondeau, et basé à Hong Kong.
- **Algorithmia** (2013, USA, $12,9M) propose son outil logiciel AI Layer pour déployer dans le cloud et de manière répartie des modèles entraînés de machine learning. Le tout est associé à une bibliothèque de plusde 4000 micro-services intégrables dans les applications. Il en résulte un modèle déployé dans le cloud exploitable sous forme de service. L'outil fait gagner du temps aux data scientists et automatise la partie « dev ops » (les opérations de déploiement des applications à la croisée des chemins entre les développeurs et les responsables de l'infrastructure IT) du déploiement du projet.
- **Skymind** (2014, USA, $6,3M) est une autre startup qui facilite le déploiement de solutions de machine learning dans le cloud avec sa Skymind Intelligence Layer (SKIL). Elle est notamment adaptée à l'entraînement incrémental de modèles de machine learning. L'outil est surtout destiné aux équipes informatiques qui déploient les applications. La startup est aussi à l'origine de Deeplearning4j.org, un framework open source de deep learning pour Java, distribué sur Apache Spark et Hadoop, écrit en C et C++ et CUDA pour l'accès aux GPU Nvidia.
- **Ayasdi** (2008, USA, $106M) interprète aussi de gros volumes de données pour y identifier des signaux faibles pertinents. Le projet a démarré à Stanford et avec des financements de la DARPA et de la NSF, l'équivalent américain de l'Agence Nationale de la Recherche française.

- **Versive** (2012, $57M), anciennement Context Relevant, propose des outils d'analyse prédictive applicables à différents marchés. Le glissement sémantique semble généralisé : au lieu de parler de big data, ce qui est trop vague, les startups parlent plutôt d'analyse prédictive qui exploite de gros volumes de données. Serait-ce de l'IA washing ? Conceptuellement oui, même si ce genre d'entreprise utilise probablement des briques de réseaux neuronaux et de machine learning en plus de méthodes plus traditionnelle.

Les applications de l'IA font aussi appel aux ressources du cloud, en particulier dans les phases d'entraînement et surtout pour les startups qui ne peuvent pas disposer de leur propre data center.

Les grandes entreprises auront à gérer un équilibre entre leurs data-center « on premise » (chez elles) et dans des clouds privés et publics. La rapidité d'évolution des technologies de processeurs neuromorphiques et GPU que nous avons vues plus haut justifie le choix du cloud pour éviter l'obsolescence rapide de ses infrastructures.

Les infrastructures en cloud doivent pouvoir «scaler » pour s'adapter à l'entraînement de modèles de machine learning et deep learning nécessitant d'aligner parfois des milliers de serveurs. Une fois les modèles entraînés, leurs besoins en ressource machine sont plus faibles, surtout pour les solutions de deep learning. Ce n'est pas pour rien, par exemple, qu'un GPU Nvidia ou un Google TPU offre une puissance de calcul située aux alentours du 100 Tflops/s tandis que les unités de traitement neuronales embarquées dans les smartphones comme le Huawei Pmate 10 et l'iPhone 8/X ont une puissance de calcul située entre 1 et 2 Tflops/s ! L'exécution d'un réseau de neurones est bien plus rapide que son entraînement !

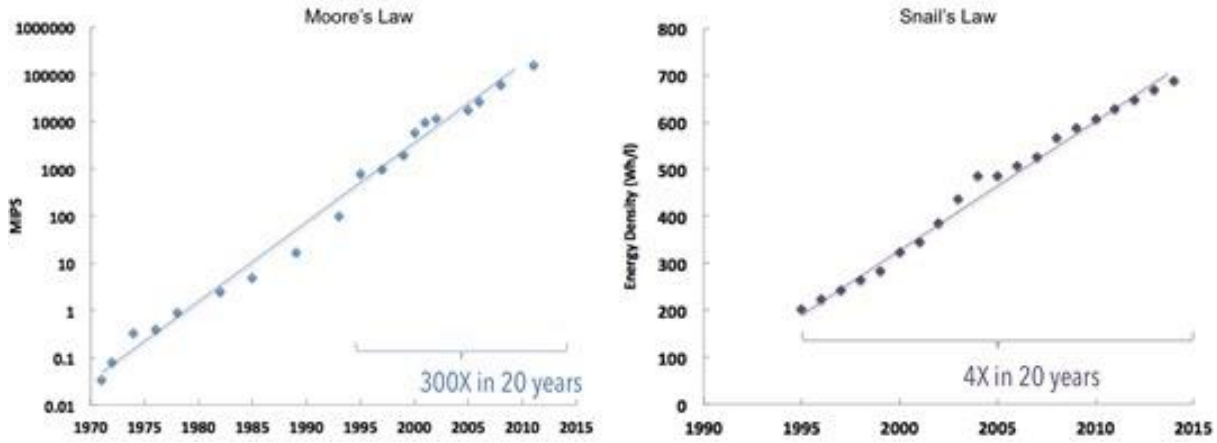
La plupart des offres de cloud intègrent maintenant la capacité à entraîner et exécuter des modèles de machine learning et de deep learning. C'est le cas chez Google, IBM, Microsoft, Amazon et même OVH qui dispose de serveurs DGX-1 de Nvidia.

Energie

L'homme ne consomme en moyenne que 100 Watts dont 20 Watts pour le cerveau. C'est un excellent rendement. Tout du moins, pour ceux qui font travailler leur cerveau. Ce n'est pas facile à égaler avec une machine et pour réaliser les tâches de base que réalise un humain. Les supercalculateurs consomment au mieux quelques KW et certains dépassent les MW.

Des progrès sont cependant notables dans les processeurs mobiles. Consommant moins de 5 W, ils agrègent une puissance de calcul de plus en plus impressionnante grâce à des architectures multi-cœurs, à un fonctionnement en basse tension, aux technologies CMOS les plus récentes comme le FinFET (transistors verticaux) et/ou FD-SOI (couche d'isolant en dioxyde de silicium réduisant les fuites de courant dans les transistors et améliorant leur rendement énergétique) et à une fréquence d'horloge raisonnable (entre 1 et 1,5 GHz). La technologie FD-SOI issue de STMicroelectronics et Soitec gagne petit à petit du terrain, notamment chez Samsung, Global Foundries et NXP. On a aussi déjà vu le rôle des memristors pour améliorer l'équation énergétique du deep learning.

La mécanique et l'énergie sont les talons d'Achille non pas de l'IA qui est distribuable là où on le souhaite mais des robots. Un homme a une autonomie d'au moins une journée en état de marche convenable sans s'alimenter. Un robot en est encore loin. D'où l'intérêt des travaux pour améliorer les batteries et notamment leur densité énergétique. Un besoin qui se fait sentir partout, des smartphones et laptops aux véhicules électriques en passant par les robots. Les progrès dans ce domaine ne sont pas du tout exponentiels. Cela a même plutôt tendance à stagner. Dans les batteries, c'est la loi de l'escargot qui s'appliquerait avec un quadruplement de la densité tous les 20 ans (**source**).



Des laboratoires de recherche inventent régulièrement des technologies de batteries battant des records en densité énergétique ou du côté du temps de chargement, à base de matériaux différents et/ou de nano-matériaux, ou de composés différents au lithium. Il y a notamment le lithium-sulfure ou le lithium-oxygène permettant en théorie d'atteindre une densité énergétique 20 fois supérieure à celle des batteries actuelles, utilisées dans les véhicules électriques[298].

Mais en elles sortent rarement, faute de pouvoir être industrialisées à un coût raisonnable ou de bien fonctionner dans la durée. Parfois, on arrive à une densité énergétique énorme, mais cela ne fonctionne que pour quelques cycles de charge/décharge. Trop injuste !

Résultat, pour le moment, la principale voie connue est celle de l'efficacité industrielle, choisie par Elon Musk dans la création de sa Gigafactory dans le Nevada, une usine à $5B qui exploitera la technologie de batteries standards de Panasonic, qui a aussi mis $1B au pot pour le financement de l'usine. Une usine qui est aussi proche d'une **mine de Lithium**, à Clayton Valley, l'un des composés clés des batteries et qui démarrera sa production en 2020.

On peut cependant citer l'étonnante performance d'un laboratoire de l'université de Columbia qui a réussi à alimenter un composant CMOS avec de l'énergie provenant de l'ATP (adénosine triphosphate), la source d'énergie principale des cellules vivantes qui est générée par les nombreuses mitochondries qu'elles contiennent. Cela ouvre des portes vers la création de solutions hybrides biologiques et informatiques insoupçonnées jusqu'à présent. La véritable bionique en quelque sorte !

[298] Voir Stockage de l'électricité: les batteries lithium-ion face au tout pétrole de Nicolas Hahn, décembre 2013.

Applications génériques de l'IA

Après avoir décrit les techniques de base de l'intelligence artificielle côté logiciels, algorithmes, données et matériel, nous allons passer aux applications et aux usages.

Cette partie est dédiée aux applications génériques de l'IA qui sont généralement transversales aux entreprises de tous secteurs d'activité.

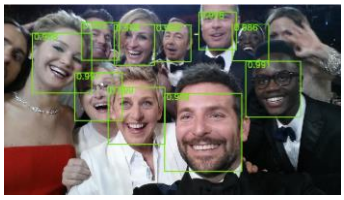
Elles sont organisées en deux grandes catégories :

- La **vision**, le **langage** et la **robotique** qui s'appuient sur les briques fondamentales de l'IA vues précédemment.
- Le **marketing**, les **ressources humaines**, la **comptabilité** et la **cybersécurité** qui sont des fonctions horizontales dans les entreprises et font appel aux briques technologiques de l'IA ainsi qu'aux trois domaines précédents, en fonction des besoins.

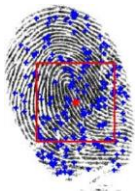
Vision

La vision artificielle est l'application la plus courante et diversifiée de l'IA. C'est l'une des principales applications du deep learning. C'est dans ce domaine que les progrès de l'IA ont été les plus manifestes ces 10 dernières années.

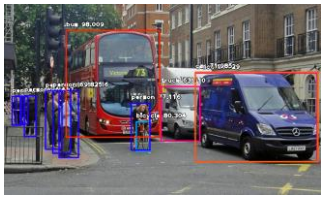
La recherche continue d'avancer dans ce créneau, en particulier dans les techniques de reconnaissance d'images pour élever au maximum le niveau sémantique de l'identifier des personnes, objets et même terrains dans le cadre de la télédétection.



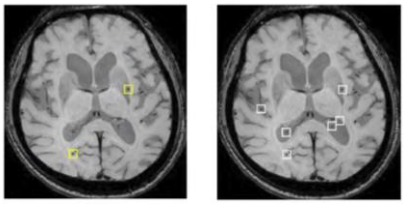
visages



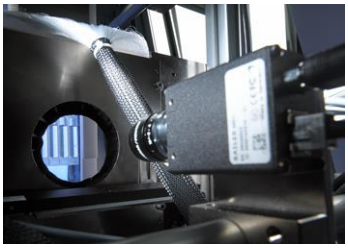
empreintes



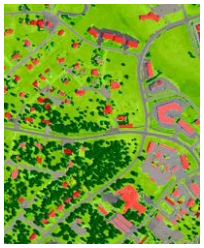
activités



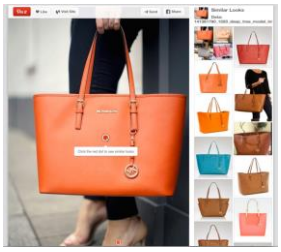
imagerie médicale



contrôle qualité



télédétection



images similaires

génération

Nous allons passer en revue les principaux usages de la vision artificielle et leurs progrès les plus récents.

Reconnaissance d'images

On la trouve pour les moteurs de recherche, les réseaux sociaux et les systèmes de sécurité et/ou vidéosurveillance. Elle est aussi utilisée couramment dans les appareils photos pour la mise au point et pour la détection des sourires (*ci-dessous* à droite).

Etat de l'art

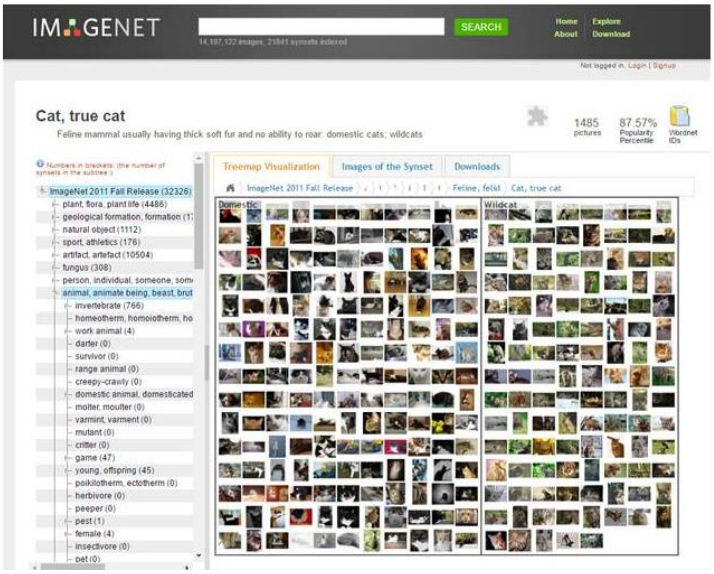
La reconnaissance et la labellisation d'images est une technologie qui progresse rapidement depuis 2012 grâce au deep learning et aux GPU.

Dans ses premières prouesses en 2012, le deep learning était capable de reconnaître un objet unique dans une image. L'état de l'art correspondant figure *à droite*, issu de ImageNet Classification with Deep Convolutional Neural Networks de Alex Krizhevsky, Ilya Sutskever et Geoff Hinton de l'Université de Toronto (9 pages). Le réseau ne neurones était entraîné sur une base de 1,2 millions d'images. Il pouvait reconnaître un millier de types d'objets différents mais avec un taux d'erreur de 37%, inacceptable pour tout système en production.

Le réseau convolutionnel comprenait cinq couches et 60 millions de paramètres répartis dans 650 000 neurones. Il était entraîné sur des GPU Nvidia GTX 580 configurés avec 3 Go de mémoire. Les images en entrées faisaient 224 pixels de côté, en couleur.

La base d'entraînement était issue de la base de référence d'images **ImageNet**.

En 2016, elle en comportait plus de 10 millions d'images, et atteignait 14,2 millions en septembre 2018[299]. Elle sert de benchmark aux solutions de reconnaissance d'images. Google et Facebook disposent de bases d'entraînement encore plus grandes, de plus de 100 millions d'images pour le premier et avec plus de 20 000 classes d'objets différentes.



Trois ans plus tard, en 2015, il devenait possible d'identifier plusieurs objets et personnes dans une même image comme dans **DenseCap**[300] qui associe un réseau convolutionnel pour la détection des objets puis leur labellisation.

Il alimente un réseau à mémoire LSTM qui génère un label en texte clair des objets détectés. Les images traitées faisaient 720 × 600 pixels.

[299] L'augmentation de la base de référence ne change rien au dimensionnement du réseau de neurones convolutionnel. Il rallonge surtout son temps d'entraînement. L'augmentation du nombre de classes d'objets complexifie le réseau dans les couches finales de neurones dites « fully connected » qui font le lien entre les dernières feature maps et les classes d'objets.

[300] Voir DenseCap: Fully Convolutional Localization Networks for Dense Captioning de Ku Fei-Fei, Stanford, 2015.

DenseCap: Fully Convolutional Localization Networks for Dense Captioning, J. Johnson, A. Karpathy, L. Fei-Fei, 2015: http://arxiv.org/abs/1511.07571

La même année, des équipes de Google étaient capables de décrire une scène comportant plusieurs personnes et un objet (un frisbee), la prouesse étant d'ailleurs plutôt située dans l'agencement du réseau à mémoire **LSTM** que dans la reconnaissance des composantes de l'image[301]. Le système avait l'air de fonctionner dans une belle diversité de situations[302]. Depuis 2017, l'application de réalité augmentée mobile **Google Lens** ajoute des informations sur les objets visés par la caméra du smartphone, comme des fleurs et des restaurants[303].

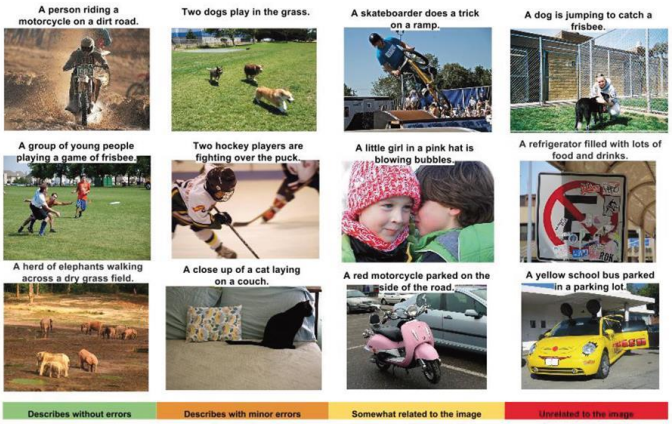


Figure 5. A selection of evaluation results, grouped by human rating.

L'état de l'art progresse grâce au défi **ImageNet Large Scale Visual Recognition Challenge** (ILSVRC) lancé en 2010 et renouvelé chaque année[304], le dernier datant de juillet 2017. Il permet d'évaluer l'état des lieux de la reconnaissance d'images en mettant en concurrence plus d'une cinquantaine d'entreprises et laboratoires de recherche dans le monde.

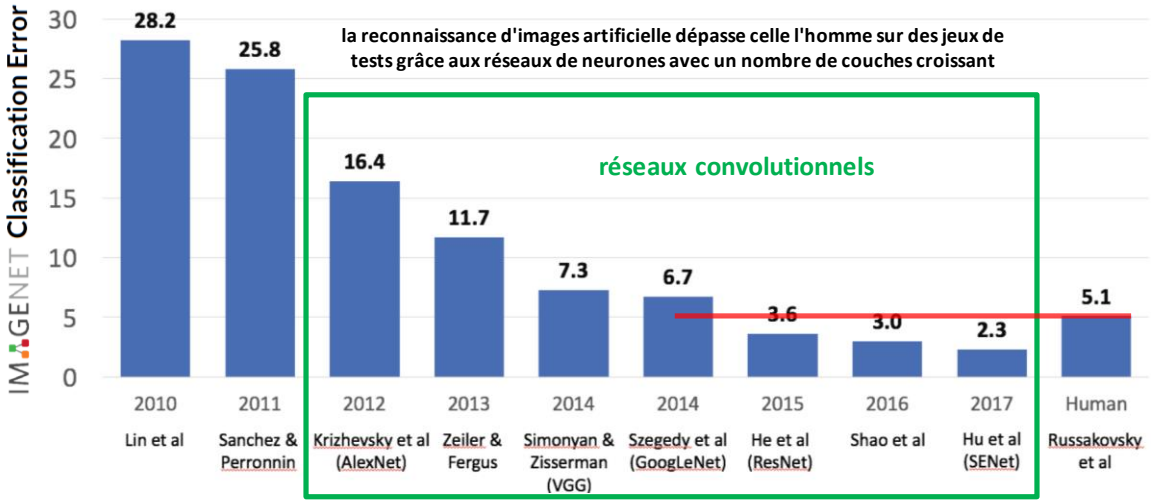
[301] Voir Show and Tell: A Neural Image Caption Generator, Google, 2015 (9 pages).

[302] Ce genre d'application peut exploiter la base Visual Genome qui est un jeu de données de 107 077 images dont les objets et régions sont labellisés avec des termes reliés dans des graphes. Voir Visual Genome Connecting Language and Vision Using Crowdsourced Dense Image Annotations, Stanford, 2016 (45 pages).

[303] Voir AI-powered Google Lens can identify types of flowers, give info about restaurants, mai 2017.

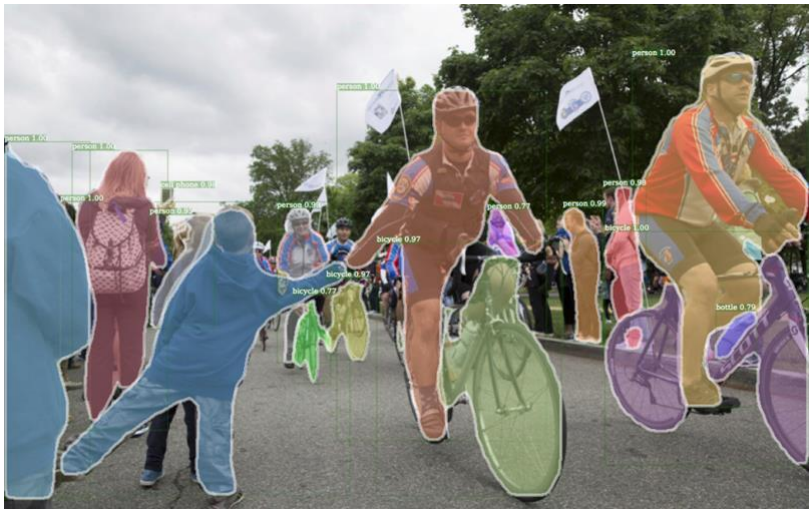
[304] Ce benchmark porte sur la reconnaissance d'images issue d'une base comprenant un million d'images dans 1000 classes différentes. Le niveau d'erreur mesuré est top-5 ou top-1. Le top-5 correspond à la proportion d'images pour lesquelles le bon label ne figure pas dans les cinq premiers considérés comme étant les plus probable par le réseau de neurones. Le top-1 correspond au label le plus probable. C'est le score le plus intéressant, le plus proche de la reconnaissance humaine. Le top-5 est un peu trop laxiste !

Les réseaux de neurones utilisés sont de plus en plus profond (nombre de couche) et de plus en plus larges (nombre de catégories d'objets reconnus et taille des bases d'entraînement). Pour la reconnaissance d'images, on peut dépasser 150 couches de neurones (mais en général, toujours avec environ cinq couches de convolution).



Lancé en open source début 2018, le projet **Detectron** de Facebook[305] associe un grand nombre de réseaux de neurones pour détecter et bien détourer les objets dans une image.

Il est développé en Python avec le framework Caffe2. Il utilise les réseaux de neurones Mask R-CNN, RetinaNet (tous deux créés en 2017), Faster R-CNN, RPN, Fast R-CNN, R-FCN et s'appuient sur ResNeXt, ResNet et un réseau convolutionnel VGG16. Cela illustre la notion d'intégration dans l'IA[306].



Pour une fonction, il faut combiner ici une dizaine de techniques différentes !

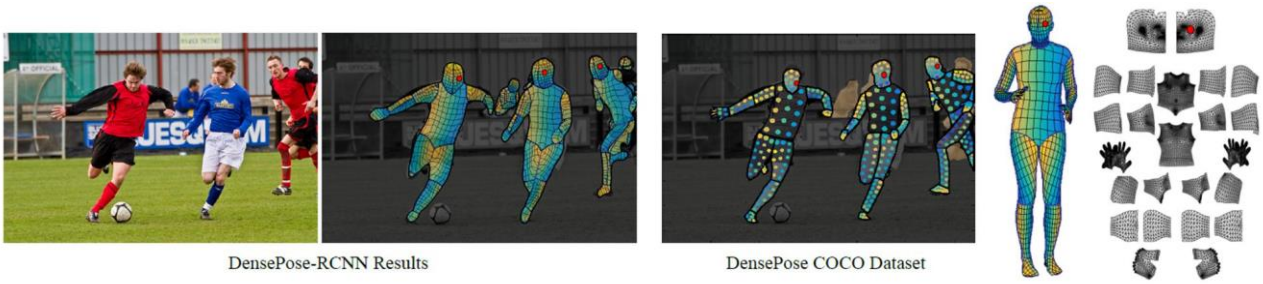
Ces techniques sont exploitées dans les systèmes de conduite assistée et autonomes comme chez **Mobileye**, filiale d'Intel depuis 2017, qui détecte les piétons, les cyclistes, les autres véhicules, la signalisation au sol et les panneaux de signalisation avec de simples caméras RGB.

En avril 2018, Google faisait encore mieux en introduisant la reconnaissance de races de chiens dans l'application mobile Google Lens (source).

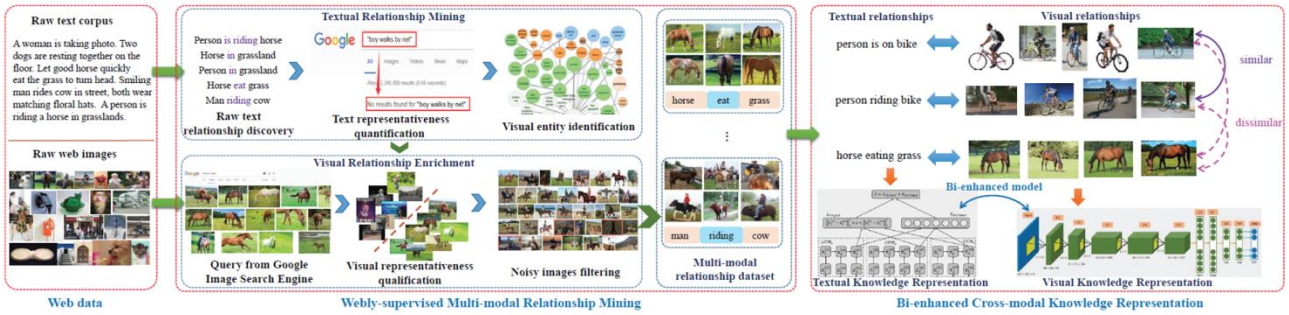
[305] Voir https://research.fb.com/downloads/detectron/. Cela fait suite à Fully Convolutional Networks for Semantic Segmentation 2016 (12 pages) issu de l'Université de Berkeley.

[306] Google propose une fonction de segmentation d'images équivalente, **DeepLab**. Elle est notamment employée dans les smartphones Pixel 2 et 2 XL pour flouter l'arrière-plan de portraits.

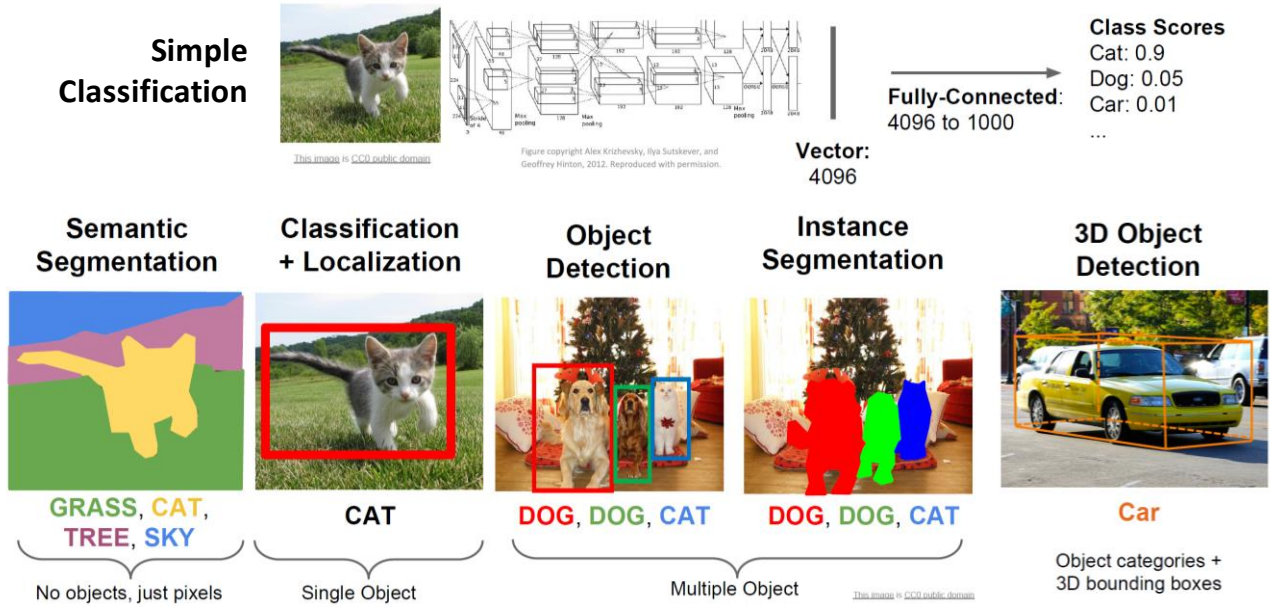
Début 2018, le projet **DensePose** réalisait une prouesse de plus en étant capable de bien isoler les différentes parties du corps dans la reconnaissance d'images de personnes (*ci-dessous*)[307].



La reconnaissance d'images peut servir à la représentation des connaissances en s'appuyant sur une approche multimodale associant des descriptifs textuels riches d'images et les images proprement dites[308].



De nombeux types de réseaux de neurones permettent aussi de reconstituer la profondeur des objets dans une image sans même disposer de vision stéréoscopique[309].



[307] Voir DensePose: Dense Human Pose Estimation In The Wild de Riza Alp Güler (INRIA-Centrale-Supelec), Natalia Neverova et Iasonas Kokkinos (Facebook), 2018 (pages).

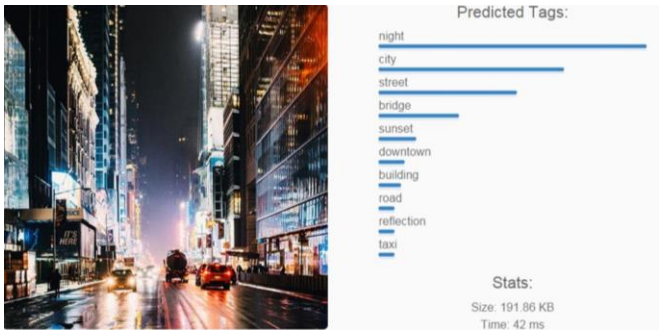
[308] Voir Multi-Modal Knowledge Representation Learning via Webly-Supervised Relationships Mining, 2018 (9 pages).

[309] Voir Deep learning for 3-D Scene Reconstruction and Modeling, 2017 (147 slides)

En résumé, au-delà de la simple classification d'une seule image dans une catégorie, nous avons la segmentation sémantique qui colorie les pixels d'une image en fonction des objets et zones détectés, la classification et la localisation qui détecte les objets et les entoure d'un cadre, avec un seul objet ou plusieurs objets, puis la segmentation d'instances d'objets différents et enfin, la détection de la forme 3D des objets qui est possible avec des images en 2D voir plusieurs images d'un même objets pris sous plusieurs angles, ou enfin, via une vue 3D obtenue par exemple avec un capteur de profondeur ou un LiDAR intégré dans une voiture[310].

Les progrès les plus récents concernent la détection des objets dans une vidéo. Elle opère maintenant en temps réel [311]!

L'interprétation des images est un pan entier de l'IA qui est la spécialité de nombreuses startups qui n'ont pas toutes été acquises par les GAFA ! Ces startups utilisent des techniques assez voisines basées sur le deep learning pour identifier le contenu de photos ou de vidéos, pour en extraire des labels qui sont ensuite exploitées dans diverses applications. En voici quelques-unes.



Clarifai (2013, USA, $40M) propose une API en cloud permettant d'accéder à leurs fonctions de reconnaissance d'images. La startup a été créée par Matthew Zeiler, un ancien de l'équipe de Jeff Dean chez Google. Ils visent des marchés divers comme le e-commerce et les médias. Leur solution intègre une fonction de recherche des les images.

Vicarious (2010, USA, $122M) est spécialisé dans la reconnaissance et la classification d'images. Ils se sont fait remarquer en étant capable d'interpréter des Captcha de toutes sortes avec une efficacité de 90%. Ils sont maintenant focalisés sur les usages de la reconnaissance d'image en robotique.

Cortica (2007, Israël, $69,4M) extrait les attributs clés d'images fixes ou animées pour les associer à des descriptifs textuels avec sa solution Image2Text.

Elle est par exemple capable de reconnaître une marque et modèle de voiture dans une vidéo ou un animal dans une photo. Cortica indique faire du deep learning non supervisé et temps réel. Le tout est protégé par 200 brevets ! La startup vise le marché de l'automobile, de la vidéosurveillance et des drones.

[310] Voir le cours CS231N de Stanford de Fei-Fei Li, Justin Johnson et Serena Yeung et en particulier la onzième partie d'où est tirée l'illustration, dans Detection and Segmentation, mai 2018 (104 slides).

[311] Voir cet état de l'art : Deep Learning for Video Classification and Captioning, février 2018 (114 pages).

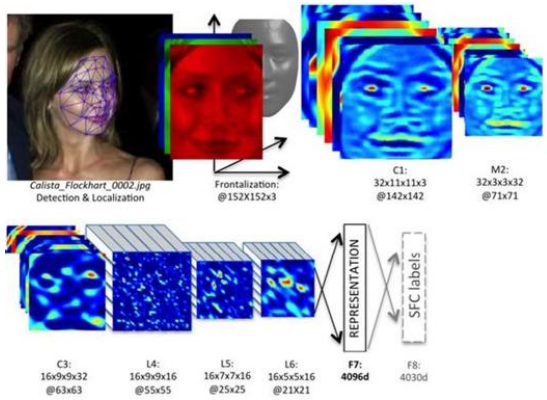
Deepomatic (2014, France, $2,3M) utilise le deep learning pour interpréter le contenu, la forme et la couleur d'images dans les médias et les associer à des publicités contextuelles. Le tout pour de nombreux marchés (santé, transports, sécurité, industrie).

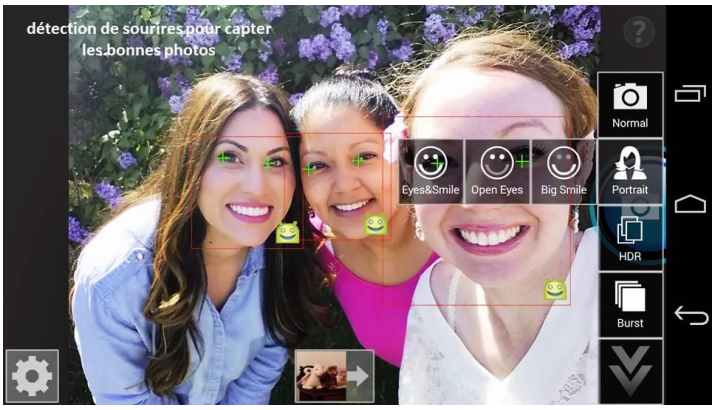
D'autres startups, notamment chinoises, se focalisent sur la détection d'activités en exploitant les images de caméras de surveillance.

C'est ce que propose notamment **Umbo CV** (2014, Taïwan, $2,8M) qui s'appuie sur une solution logicielle fonctionnant dans le cloud.

Visages

La reconnaissance de visages est une technologie plutôt au point et depuis quelques années. Dans leur projet FaceNet, **Google** annonçait en 2015 avoir atteint un taux de réussite de détection de visage de 99,63%[312].





De son côté, **Facebook** et son projet DeepFace s'appuyait sur la technologie issue d'une start-up israélienne **face.com** avec un taux de réussite de 97,25% pour vérifier qu'une personne sur une photo est la même sur une autre, quel que soit l'angle de la prise de vue et l'éclairage. C'est très voisin du taux de reconnaissance humain qui serait évalué à 97,5%.

On trouve de la détection de visages dans plein de solutions du marché comme avec la fonction Faces d'**Apple** iPhoto[313].

Les APIs en cloud des Cognitive Services de **Microsoft** apportent des services équivalents aux développeurs d'applications. Il en va de même pour **IBM Watson**, pour **Google** avec ses Cloud Vision APIs et **Amazon** avec son SDK Rekognition. Cette abondance des offres rappelle que les technologies de l'IA, une fois au point, deviennent rapidement des commodités. Les méthodes sont sur la place publique et quasiment des commodités. Il faut ensuite les mettre en œuvre avec du logiciel et du matériel.

login par reconnaissance de visages avec Windows Home et Intel RealSense

[312] Voir FaceNet: A Unified Embedding for Face Recognition and Clustering, juin 2015 (10 pages). Le tout s'appuyait sur un réseau neuronal à 22 couches avec comme entrée, des images de 220 pixels de côté.

[313] Elle provient peut-être de la start-up suédoise **Polar Rose** acquise par Apple en 2010.

La différence se situe dans la mise en œuvre et aussi dans le marketing.

La reconnaissance des visages est évidemment un sujet chaud pour les services de sécurité. On en voit dans tous les films et séries TV ! En quelques secondes, les suspects sont identifiés. Est-ce comme cela dans la vraie vie ? Probablement pas. Cela explique pourquoi le **FBI** a lancé son projet NGI (Next Generation Identification) en 2009 qui est maintenant opérationnel. Il était pourvu à hauteur de la bagatelle de $1B et réalisé par **Lockheed Martin**.

Le marché de la reconnaissance faciale est aussi prolixe en solutions diffusées en OEM, comme **imagga** (2008, Bulgarie, $642K) et ses API en cloud de tagging automatique d'images en fonction de leur contenu et **Cognitec (**2002, Allemagne) qui vise surtout les marchés de la sécurité.

Nous avons aussi la startup française **Smart Me Up** (2012, France, 3M€, acquise par l'Italien Magneti Marelli en août 2018), qui propose une solution logicielle d'analyse des visages. Elle détecte l'âge, le comportement et les émotions des utilisateurs. La solution est commercialisée sous forme de brique logicielle en marque blanche utilisable dans des applications métier.

SenseTime (2014, Chine, $1,6B) commercialise une solution de reconnaissance de visages déclinée dans plusieurs verticaux dont le retail et les télécoms. Elle a été fondée par des chercheurs de Hong Kong.

Ses primitives fonctionnelles sont le suivi de plusieurs visages en temps réel dans des vidéos et la détection d'attributs divers (sourire, style de coiffure et barbe, age, race, regroupements pour l'organisation d'albums photos, détection de visage vivant vs statique, maquillage virtuel). Elles visent le florissant marché de la vidéosurveillance en Chine, dans la distribution mais aussi l'imagerie médicale, la conduite autonome (avec Honda).

SenseTime est partenaire du MIT dans le programme de recherche Intelligence Quest et a Qualcomm ainsi qu'Alibaba parmis ses actionnaires. C'est l'une des startups de l'IA les mieux financées au monde. Ils disposent de leur propre centre de calcul avec 8000 GPU qui supporterait des réseaux de neurones avec des centaines de milliards de paramètres.

Megvii (2011, Chine, $607M) est un concurrent local de SenseTime avec son produit Face++ de détection de visage. Ces petits joueurs n'ont levé que $607M[314]. Hahaha ! Et ne parlons pas de **CloudWalk Technology** (2006, Chine, $389M) qui cible surtout la surveillance dans les aéroports. Quand à **FaceAll** (2015, Chine), ils fournissent un SDK d'outils de reconnaissance de visages exploitable par des tiers et qui bat des records de vérification de visages (99,67% de résultats corrects).

On trouve des solutions de reconnaissance de visage dans les vidéos chez **Kairos** (2012, USA, $3,5M) qui savent aussi analyser les émotions et quantifier les foules, chez **KeyLemon** (2008, Suisse, $1,5M) qui propose une solution en cloud, chez **Matroid** (2016, USA, $3,5M), qui fonctionne sur des flux vidéo ou des photos ou chez le japonais **NEC**. Il faut aussi citer **OpenCV**, une solution open source de détection de visages et cela ne doit pas être la seule.

[314] Megvii a recruté l'un des créateurs du réseau de neurones ResNet de Microsoft Research, Jian Sun, qui est directeur scientifique de la startup depuis 2016.

La reconnaissance de visages sert évidemment aussi aux applications de vidéo-surveillance, comme celle de **Camio** (2013, USA) qui fournit une solution en cloud d'exploitation de vidéos de caméras de surveillance.

Emotions

L'analyse des visages, couplée ou non avec celle de la voix, permet de détecter des éléments extérieurs d'émotions humaines.

Le principe de la reconnaissance des émotions dans le visage à partir de caméras est assez ancienne. Elle est standardisée par le système de description **FACS** pour Facial Action Coding System[315], créé en 1978 par les psychologues américains Paul Ekman et Wallace Friesenen.

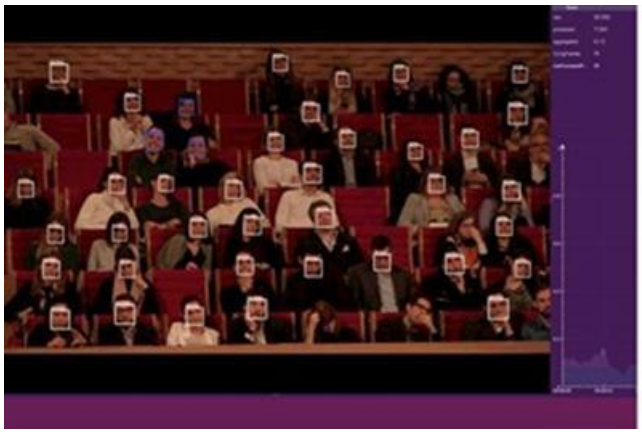
Diverses startups se sont attaquées à cette fonction pour toucher divers marchés comme ceux du commerce et de la publicité mais aussi de sécurité.

L'une d'entre elles est la startup **Affectiva** (2009, USA, $26,3M) que j'avais découverte au CES 2013[316]. Elle présentait une solution de captation les émotions d'un utilisateur exploitant une simple webcam sur un micro-ordinateur. Elle valorise un projet de recherche du MIT Media Lab et vise le marché de la publicité et du retail mais a eu du mal à le pénétrer. Ils s'intéressent depuis au marché de l'automobile pour détecter l'état du conducteur, comme le manque d'attention, qui est une fonctionnalité finalement assez limitée. Ils annoncaient avoir entraîné leurs modèles de deep learning avec 5 millions de visages issus de 75 pays.

Leur logiciel évalue les paramètres suivants : joie, gaieté, dégoût, mépris, peur, surprise, colère ainsi que la valence (allant du négatif au positif), l'engagement et l'attention. Le tout exploite l'analyse de 20 expressions faciales différentes via des réseaux convolutifs (ou convolutionnels) et des SVM. Mais il est difficile de savoir où cette solution est déployée d'un point de vue pratique.

En France, la société **Datakalab** analyse simultanément plusieurs visages, comme les spectateurs d'un événement ou d'une conférence. Sa solution peut ainsi déterminer l'intérêt d'une audience pour une présentation, ou comparer cet intérêt entre deux intervenants comme en mai 2017 pendant le débat d'entre deux tours confrontant Emmanuel Macron à Marine Le Pen. Cela objectivisait une impression partagée sur la performance relative des deux finalistes ! L'opération était aussi réalisée sur un bout de l'audience de l'Echappée Volée début juillet 2018 à la Seine Musicale de Boulogne Billancourt pour comparer l'impact des différents intervenants.

Le service permet aussi d'évaluer le niveau de stress de clients, comme dans l'arrivée en gare. Datakalab se positionne comme un cabinet de conseil en neuromarketing. Il n'exploite pas que la vidéo mais aussi les informations issues de la voix et, optionnellement, de bracelets biométriques.



L'analyse des gestes et autres mouvements est un autre domaine où l'IA peut jouer un rôle. Il est pour l'instant moins courant que l'analyse des visages, mais se développe de plus en plus. Côté recherche, l'étude européenne **Survey on Emotional Body Gesture Recognition**, publiée en janvier 2018 (19 pages) fait un état des lieux. Elle illustre le fait que ce domaine est encore nouveau.

---

[315] Ici : https://www.paulekman.com/wp-content/uploads/2013/07/Facial-Sign-Of-Emotional-Experience.pdf.

[316] Voir le Rapport CES 2013, page 256.

L'équipe de recherche a mené une expérience avec un système à base de caméra et de Microsoft Kinect pour classifier des gestes et identifier les émotions associées. Elle met en avant le fait que la signification de ces émotions dépend de nombreux paramètres comme la culture des individus ainsi que le genre de la personne observée.

Encore plus surprenants, ces travaux de recherche publiés dans DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition using Low-Cost Thermal Imaging in Unconstrained Settings en aout 2017, visent à détecter le niveau de stress d'utilisateur avec des caméras infrarouges analysant le rythme de respiration des humains avoisinants. Le tout exploite un simple réseau de neurones convolutionnel !

Enfin, Apple faisait l'acquisition, début 2016, de la startup **Emotient**, spécialisée dans la reconnaissance d'émotions faciales à base de machine learning. Le matching de visages est une chose, mais détecter les émotions en est une autre et on peut s'attendre à ce qu'Apple utilise cette fonctionnalité dans les évolutions de ses solutions, notamment dans la visioconférence Facetime.

Imagerie médicale

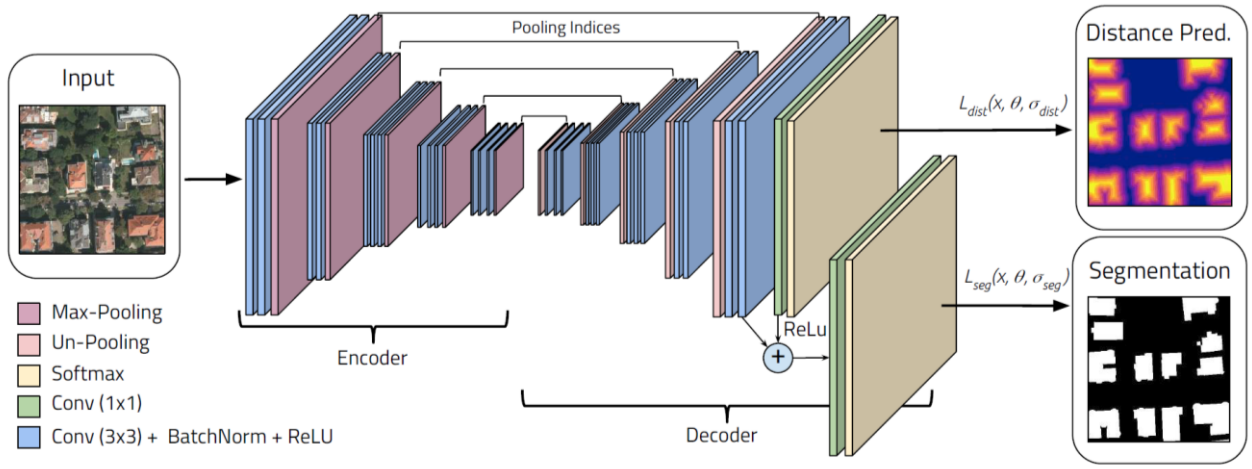
L'imagerie médicale est l'une des plus grosses applications du traitement d'images avec le deep learning. Elle est appliquée à la détection automatisée d'un grand nombre de pathologies au niveau de l'œil, de l'oreille, en dermatologie et en cancérologie. Elle exploite de nombreux réseaux de neurones différents en fonction des caractéristiques des images à traiter. Nous verrons cela plus en détail dans la rubrique sur la santé.

Télédétection

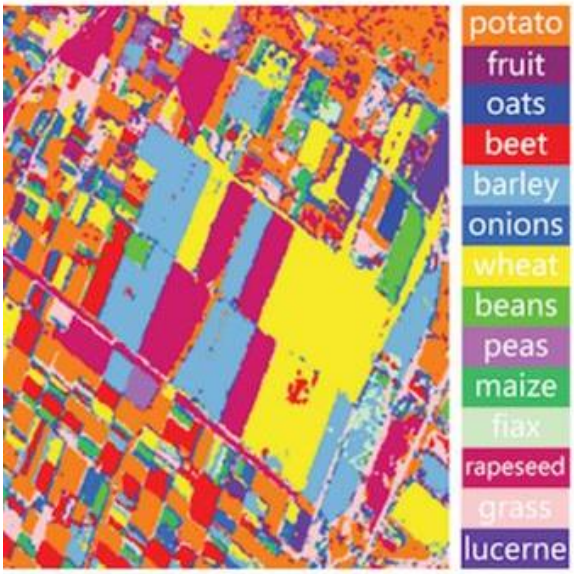
La reconnaissance d'images a aussi des applications en télédétection et imagerie satellite. Le deep learning permet de créer des solutions de recherche sémantique dans de gros volumes d'images, pour détecter des objets spécifiques comme des champs agricoles, des panneaux solaires ou des éoliennes, des plans d'eaux, des forêts, des toits, pour les caractériser en fonction de leur spectre lumineux, et pour analyser des variations dans le temps de ces paramètres.

Les innovations dans ce domaine comme ailleurs dans l'IA démarrent presque toujours au niveau de la recherche.

Ces différentes solutions utilisent une grande variété de réseaux de neurones. Ainsi, pour détecter la distance entre bâtiments et leur forme, on utilise un double *stacked auto-encoder* (*ci-dessous*).



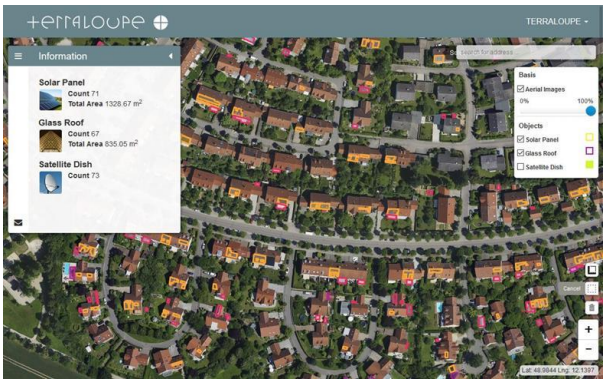
Les applications touchent la détection d'objets en biais[317]. le décompte de véhicules[318]. l'analyse de dommages sur des bâtiments, en particulier lors de conflits[319], l'analyse de dommages sur les toits après des intempéries[320], la détection de création de bâtiments dans des pays fermés comme l'Iran ou la Corée du Nord, le décompte de navires dans les ports et leurs mouvements et l'usage de données multi-spectrales couvrant le visible mais aussi l'infrarouge voir l'ultra-violet pour détecter le type de végétaux cultivés avec précision[321]. S'y ajoutent la préparation de la réponse face aux catastrophes naturelles, la détection d'inondations, l'analyse de deforestation [322], de l'impact d'incendies de forêts,[323] ou de la pollution atmosphérique.



Les applications commerciales sont aussi abondantes. En voici quelques-unes.

Airbus Defense and Space utilise le machine learning pour détecter les avions dans les aéroports avec de l'imagerie satellite. Ils les repèrent sur la base d'un jeu de données d'entraînement de 40 000 prises de vues avec la capacité à se débarrasser des nuages grâce à le mise en correspondance de plusieurs photos. Le taux d'erreur est inférieur à 4% (source).

Descartes Labs (2014, USA, $38,3M) exploite les données d'image satellite pour y découvrir comment évolue la production agricole, le cadastre des villes ou autres données géographiques, le tout via du machine learning développé sur TensorFlow et déployé sur Google Cloud. Ils prédisent la production agricole à l'échelle mondiale ainsi que les risques de famine dans les pays émergents ! Voir cette vidéo montrant l'évolution dans le temps de la végétation aux USA et un exemple *ci-dessous à gauche* de détection automatique d'éoliennes (issu de cette vidéo).



[317] Vue dans DOTA: A Large-scale Dataset for Object Detection in Aerial Images, 2018 (17 pages).

[318] Voir Real-Time Ground Vehicle Detection in Aerial Infrared Imagery Based on Convolutional Neural Network, 2018 (19 pages).

[319] Voir Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach, 2018 (8 pages).

[320] Voir Roof Damage Assessment using Deep Learning, 2018 (6 pages).

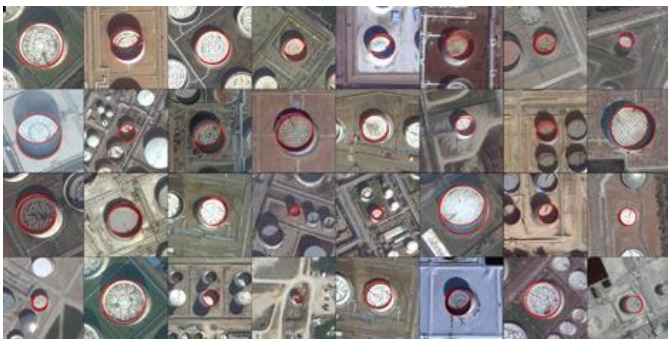
[321] Vu dans Deep Learning in Remote Sensing: A Review, 2018 (60 pages).

[322] Voir Emergency Response with Deep Neural Networks and Satellite Imagery, 2017 (76 slides) et Monitoring Deforestation in Rainforests Using Satellite Data A Pilot Study from Kalimantan, Indonesia, 2018 (26 pages).

[323] Voir cette thèse Making the most of machine learning and freely available datasets: A deforestation case study de Helen Mayfield, 2015 (368 pages).

TerraLoupe (2015, Allemagne, $788K) analyse aussi les images satellite pour reconnaître ce qu'elles contiennent, en fonction des besoins clients (vidéo), comme analyser la surface des bâtiments dans le foncier, le type de toit, les antennes satellites, les panneaux solaires avec des applications dans l'agriculture, l'immobilier ou l'assurance (exemple *ci-dessus à droite*).

C'est aussi l'activité de **Cape Analytics** (2014, USA, $31M) et d'**Orbital Insights** (2013, USA, $78,7M). Ce dernier décline son offre sur divers marchés, dont celui de l'énergie, pour évaluer les stocks et la consommation d'énergies fossiles en observant le niveau de profondeur des cuves de stockage [324] (*ci-contre*), ou celui du retail pour identifier le trafic automobile dans les centres commerciaux et optimiser l'emplacement de nouveaux points de vente.

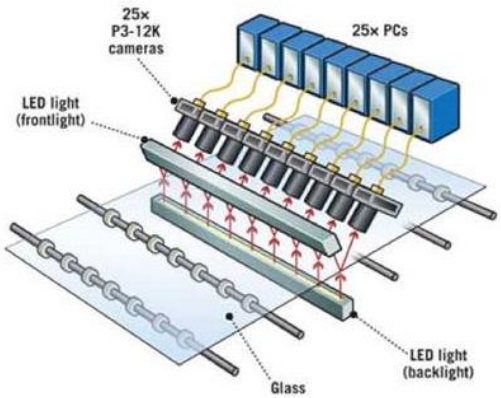


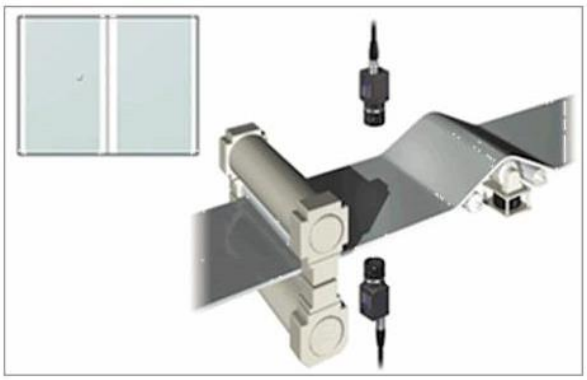
Contrôle qualité

Le contrôle qualité est très courant et se démocratise pour vérifier la qualité des pièces et produits fabriqués en usine. Il l'est aussi comme nous le verrons plus loin dans l'agro-alimentaire.

Le point clé de ces systèmes est qu'il doivent fonctionner en temps réel. Mais leur apprentissage est moins complexe car ils doivent analyser des images dont la variance est plutôt faible. Le contrôle qualité utilisait le machine learning avant l'avènement des réseaux convolutionnels. Mais ces derniers ont permis de mettre en place plus facilement des solutions avec un entraînement automatique de réseaux de neurones.

Les solutions de contrôle qualité s'appuient sur l'exploitation d'imagerie dans le visible, l'infrarouge, les UV et même les rayons X. L'imagerie peut-être complétée par d'autres types de capteurs comme ceux qui mesurent la variation de la résistance électrique de matériaux. Tous ces systèmes doivent fonctionner en temps réel, au rythme de la fabrication dans les usines, d'où l'intérêt d'utiliser des caméras intégrant des processeurs neuromorphiques exploitant des réseaux de neurones déjà entraînés. Voir d'autres solutions du genre dans la partie fabrication et contrôle qualité du vertical industrie.





Qualité d'images

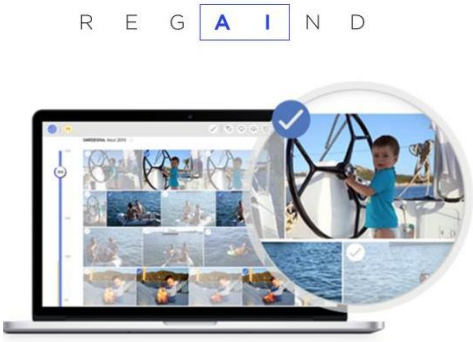
Des réseaux de neurones peuvent aussi servir à détecter la qualité d'images. C'est ce que fait **Imginit** (2017, France), en analysant la qualité d'images, pour l'acquisition d'images tierces parties, notamment dans les médias, la publicité, la gestion de stocks de photos.

---

[324] Voir https://orbitalinsight.com/products/energy/#slider-5.

Regaind (2014, France, 400K€) est une startup qui proposait une solution de tri automatique de photos en cloud s'appuyant sur du diverses méthodes de machine learning et de deep learning.

Elle permettait de trier les photos sous un angle à la fois narratif et descriptif et de les tagger automatiquement. Elle comparait diverses caractéristiques des photos : leur cadrage, le flou d'arrière plan, les couleurs, etc. La startup a été acquise par Apple pendant l'été 2017 et depuis, c'est le trou noir.



tri qualitatif automatique de photos pour créer des albums

racheté par une société US en 2017

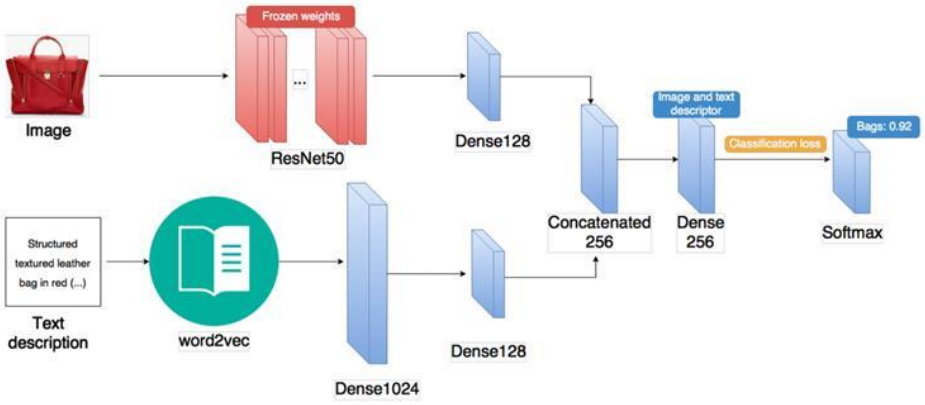
La fonction devrait logiquement faire son apparition dans des logiciels mobiles ou pour MacOS. **Google** Photo propose aussi une fonction équivalente de tri de photos.

Vous trouverez d'autres exemples d'usages de l'IA dans le secteur de la photo dans la rubrique associée dans le marché des médias et des contenus.

Recherche

Les moteurs de recherche sont de gros utilisateurs de machine learning et en particulier pour l'indexation et la recherche d'images. La création de métadonnées d'images provient à la fois des données des pages web qui les contiennent et de leurs noms de fichiers, mais elles peuvent être enrichies par la reconnaissance d'objets à base de deep learning.

Un moteur de recherche d'objets peut retrouver leurs images à partir de leur description textuelle en associant plusieurs types de réseaux de neurones (*ci-contre*) [325].



En juillet 2018, **Microsoft** lançait une fonction de recherche similaire à celle de Google Lens dans Bing.

Elle permet de trouver un objet dans les sites de vente en ligne à partir d'une photo captée par son smartphone[326].

Quelques startups proposent des solutions de recherche d'images pour les entreprises. Mais ces fonctions sont largement utilisées par les grands acteurs du Web et du mobile manipulant des images, tels **Facebook**, **Instagram**, **Pinterest**[327], **eBay**[328] et **Amazon**.

La recherche d'images similaires est aussi courante. Elle sert à la recommandation dans les systèmes de vente en ligne, dans les moteurs de recherche ainsi qu'à l'identification de contrefaçons de produits de marques. Elle s'appuie aussi sur du deep learning.

---

[325] Voir DeepStyle: Multimodal Search Engine for Fashion and Interior Design, janvier 2018 (11 pages).

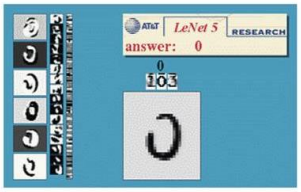
[326] Voir Microsoft launches AI-powered Bing Visual Search, juillet 2018 (vidéo).

[327] Voir Visual Search at Pinterest, 2015 (10 pages).

[328] Voir Visual Search at eBay, 2017 (10 pages) qui évoque l'utilisation d'un Convnet ResNet pour la classification d'images.

Caractères et écriture

Elle est réalisée dans les textes (OCR) issus de scans. Les systèmes actuels savent détecter les textes, les images et schémas de documents scannés. Nous avions même un leader en France dans le domaine avec la société **LTU**, acquise par le japonais **Jastec** en 2005.



Yann Le Cun !

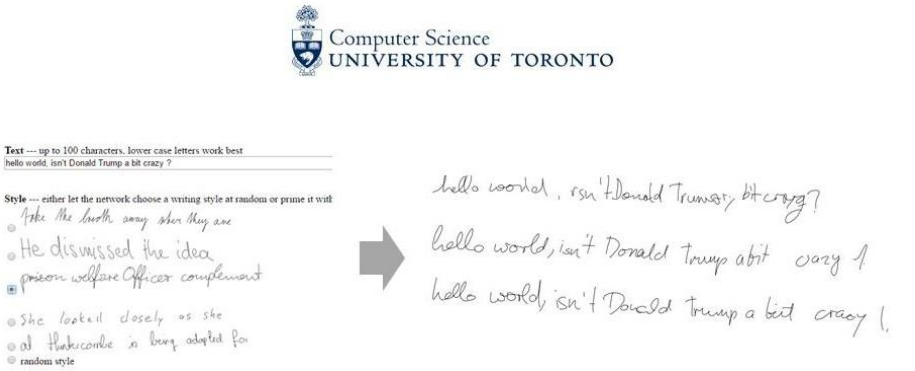
reconnaissance de caractères (OCR)

La reconnaissance de l'écriture manuscrite à partir d'encre digitale, saisie par exemple avec un stylet comme sur les tablettes.

Ce marché est moins connu que pour la reconnaissance vocale ou d'images. Et nous y avons un champion français avec la société **MyScript** (1998, France), anciennement Vision Objects, qui est basée à Nantes et qui a notamment vendu son logiciel à **Samsung**.

Nous avons aussi le californien **Captricity** (2011, USA, $52M), qui extrait les informations de l'écriture manuscrite et convertit automatiquement les formulaires en tableaux avec des applications évidentes dans les assurances et toutes les bureaucraties imaginables (vidéo). La solution en cloud est même reliée à SalesForce.

Ces systèmes sont d'ailleurs réversibles car on peut aussi générer de l'écriture manuscrite synthétique à partir d'une écriture existante comme dans ce projet de recherche de l'Université de Toronto (*ci-contre*). C'est moins connu que les voix synthétiques.



Réseaux de neurones génératifs

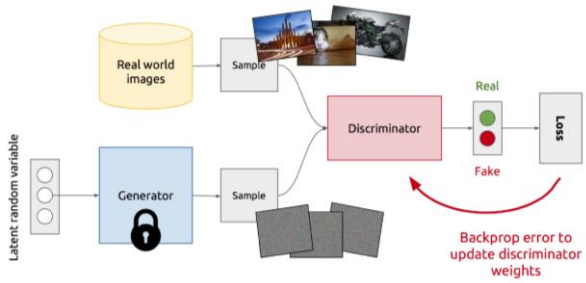
Les réseaux de neurones générateurs de contenus sont apparus en 2014 avec la publication de Generative Adversarial Nets (9 pages) par Ian Goodfellow de l'Université de Montréal avec une équipe comprenant aussi Yoshua Bengio.

Ces « GAN » impressionnent par leurs capacités à « prédire » l'univers visuel à partir de peu d'informations. Ils complètent des images ou les transforment et génèrent des images assez plausibles pour le cerveau humain. Au point que l'on en vient à trouver que l'IA est créative. Mais elle ne fait dans ces cas qu'appliquer des algorithmes de la même manière et sans discernement, comme n'importe quel filtre de retouche d'images dans Photoshop. Et elle exploite des éléments de créativité d'origine humaine. Faut-il donc lui coller des attributs de créativité comme celle des artistes ? Pas encore, même si certains croient que ce temps est déjà venu[329].

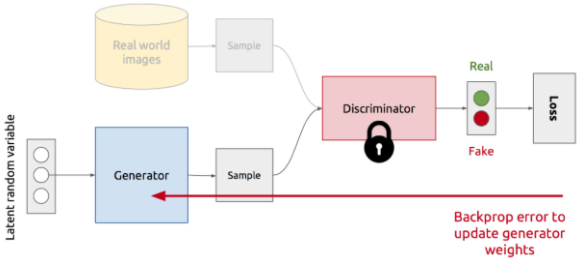
[329] Voir The Coming Creativity Explosion Belongs to the Machines, de Melba Kurman, octobre 2017 qui confond comme c'est courant la créativité des machines et celle des hommes qui les ont programmées.

Au passage, les GANs peuvent servir à générer d'autres types de contenus et notamment du langage, ce que nous verrons dans la partie associée. D'un point de vue pratique, les GANs associent un générateur de contenu et un discriminateur. Le générateur créé un contenu artificiel extrapolé à partir de contenus existants et le discriminateur vérifie que le contenu est crédible vis-à-vis d'un jeu d'entraînement.

étape 1 : entraînement du discriminateur pour qu'il reconnaisse de vraies images



étape 2 : entraînement du générateur pour qu'il génère des images reconnues par le discriminateur



Le générateur est un réseau de neurones convolutionnels de type stacked autoencoder qui génère un contenu à partir d'un autre contenu, éventuellement moins dense en information[330]. On ajuste son fonctionnement avec des paramètres divers selon les cas. Le générateur et le discriminateur se concurrencent l'un et l'autre pour ajuster le contenu généré pour qu'il soit bien reconnu comme « crédible » par le discriminateur.

Il existe en fait une très grande variété de GANs. Avinash Hindupur en a répertorié plus de 420 dans The GAN Zoo, qui était actualisé en février 2018. On y trouve par exemple les VAE (Variational Autoencoder), les DCGAN et WGAN (Wasserstein Generative Adversarial Network). Comme tous les réseaux de neurones, les GANs sont des outils probabilistes. Ils ne fournissent pas « la réponse » ou « une réponse parfaite », mais une réponse plausible parmi d'autres, et dont la plausibilité est sujette à caution.

L'entraînement des GANs est difficile. On doit disposer de bons jeux de données pour les alimenter[331]. La variété des GANs s'explique par le fait qu'ils réalisent des fonctions très spécialisées, au coup par coup. Il n'existe pas de GAN généralisé capable de tout faire. C'est un problème homothétique avec celui du Graal de l'AGI, l'intelligence artificielle généralisée.

Les exemples de GANs abondent et nous allons faire ici le tour de quelques-uns des plus marquants d'entre eux[332] . Nous verrons au passage que les travaux de recherche dans les GANs sont vite récupérés par les startups ou les grands du numérique comme Google.

Les GANs ont aussi un usage intéressant : ils peuvent aussi servir à générer des jeux de données pour entraîner des réseaux de neurones de reconnaissance d'images, notamment dans l'imagerie médicale[333] !

---

[330] Voir Generative Adversarial Networks de Ole-Johan Skrede de l'Université d'Oslo, mai 2018 (88 slides) ainsi que Generative Adversarial Networks (GANs) de Binglin, Shashank et Bhargav, 2017 (73 slides) qui expliquent bien le procédé. Cette dernière source est celle des schémas de cette page.

[331] Voir Limitations of Encoder-Decoder GAN architectures, de Sanjeev Arora et Andrej Risteski, mars 2018, qui décrit bien les limitations des GANs et leur papier Do GANs learn the distribution ? Some theory and empirics, 2018 (16 pages). Vous avez aussi quelques bon exemples de ratés dans This AI is bad at drawing but will try anyways, août 2018.

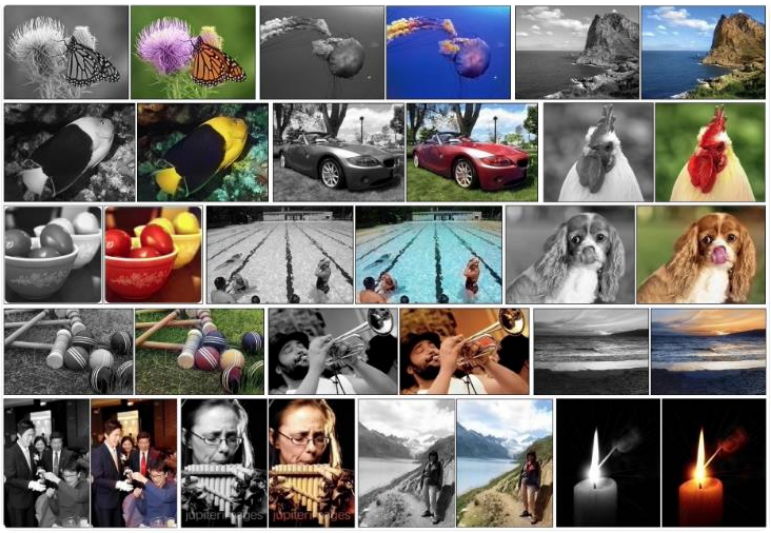
[332] Certains d'entre eux sont issus de cette présentation : Generative Adversarial Network (GAN), de Hongsheng Li (non daté, 88 slides).

[333] Voir par exemple Training artificial intelligence with artificial X-rays, juillet 2018.

Coloriage

Il existe des GANs qui savent automatiquement colorier des images en noir et blanc. Cela peut servir à moderniser des contenus anciens, souvent argentiques, aussi bien photo que cinéma.

Il faut évidemment de grandes bases d'entraînement pour que les couleurs soient les bonnes. Les exemples présentés dans les publications scientifiques sont toujours plausibles mais il doit sûrement y en avoir qui ne fonctionnent pas bien lorsqu'il peut y avoir ambiguité sur une couleur. Un ciel bleu est en fait peut-être gris, même si la couleur du ciel peut avoir un impact sur le reste d'une photo de paysage.

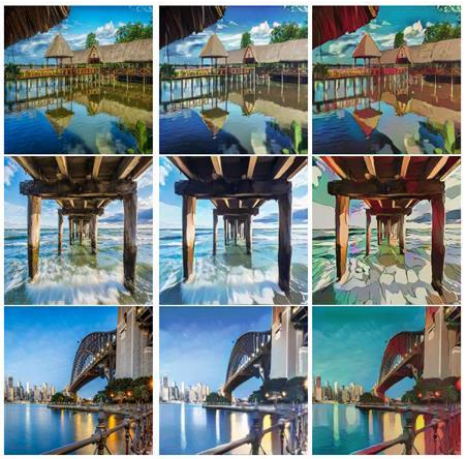


Colorful Image Colorization, Zhang, Isola & Efros, 2016

Google Photos, mai 2018

Plusieurs GANs ont été publiés pour la colorisation, comme Colorful Image Colorization de Richard Zhang, Phillip Isola et Alexei A. Efros de Berkeley 2016 (29 pages). En mai 2018, Google annonçait l'intégration de cette fonctionnalité dans son logiciel mobile Google Photos lors de la conférence Google I/O, parmis d'autres qui font aussi appel au deep learning (vidéo) comme la colorisation sélective d'images !

Dans le même registre, les GANs peuvent aussi servir à transformer des photos en bande dessinée, sans les légendes humoristiques pour l'instant. En voici un exemple récent dans CartoonGAN: Generative Adversarial Networks for Photo Cartoonization, 2018 (10 pages), réalisé en Chine (*ci-dessous*, avec les architectures en couche de son discriminateur et de son générateur).



(a) input photo (b) Shinkai style (c) Hayao style

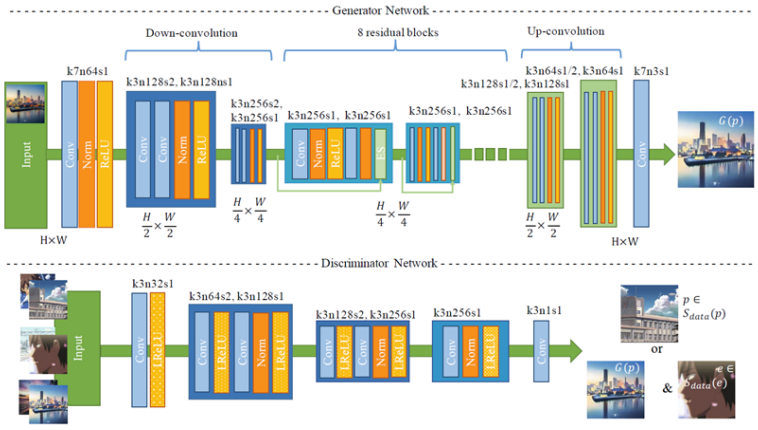


Figure 2. Architecture of the generator and discriminator networks in the proposed CartoonGAN, in which k is the kernel size, n is the number of feature maps and s is the stride in each convolutional layer, 'norm' indicates a normalization layer and 'ES' indicates elementwise sum.

Amélioration

Les GANS et/ou les stacked autoencoders permettent aussi d'améliorer des images, notamment celles qui sont bruités. C'est ce qu'a notamment démontré Nvidia en juillet en publiant Noise2Noise: Learning Image Restoration without Clean Data, 2018 (12 pages), réalisé avec des chercheurs de l'Université de Aalto en Finlande et du MIT aux USA.

Leur solution utilise des GPU Nvidia Tesla P100 et un logiciel développé avec le framework de réseau de neurones cuDNN exploitant le jeu d'instruction CUDA des GPU Nvidia. Voir la vidéo associée qui contient diverses démonstrations. Le système a été entraîné avec 50 000 images de la base ImageNet.

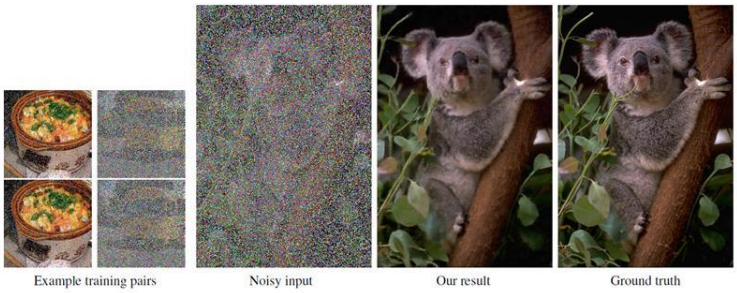


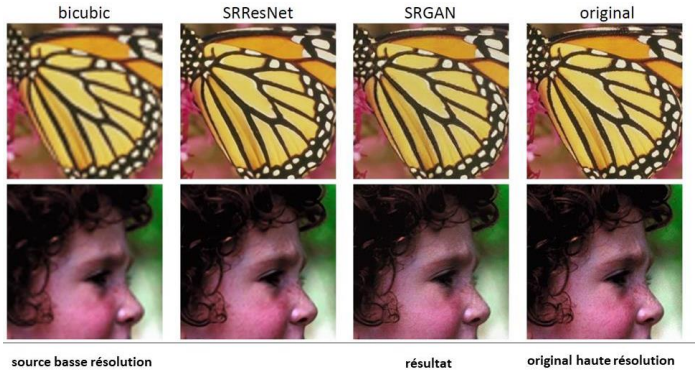
Figure 2. Random impulse noise. Our denoiser is trained on corrupted image pairs only.

Cela pourrait servir dans l'imagerie médicale et en astronomie. Mais comme l'indique un commentaire dans un article de Dpreview[334], les jeux de tests utilisés étaient à basse résolution et consommaient énormément de ressources machine. La généralisation d'un tel procédé sur des photos de reflex devra donc attendre quelques cycles de la loi de Moore !

Résolution

L'amélioration de la résolution d'images génère aussi des résultats étonnants[335]. L'idée aurait été proposée la première fois par Ian Goodfellow en 2016.

Elle est depuis déclinée dans de nombreuses productions de chercheurs qui veulent rendre leurs méthodes les plus génériques possibles. On en trouve dans Photo-realistic single image super-resolution using a generative adversarial network, de Christian Ledig et son équipe, 2016 (19 pages) qui est cité dans NIPS 2016 Tutorial: Generative Adversarial Networks, 2016 (57 pages).



Il en existe de nombreuses variantes avec A Fully Progressive Approach to Single-Image Super-Resolution (10 pages), EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis, 2017 (19 pages) qui améliore la texture des images et Learning to Super-Resolve Blurry Face and Text Images (10 pages) qui part d'une photo vraiment très floue pour générer un visage proche de la vérité (*ci-dessous*). Magique ? Non, ce sont juste des modèles probabilistes.

Figure 1. Low-resolution blurry images are challenging for the state-of-the-art super-resolution and deblurring methods. Sequentially applying super-resolution and deblurring methods further exacerbates the artifacts. Our method learns to reconstruct realistic results with clear structures and fine details.

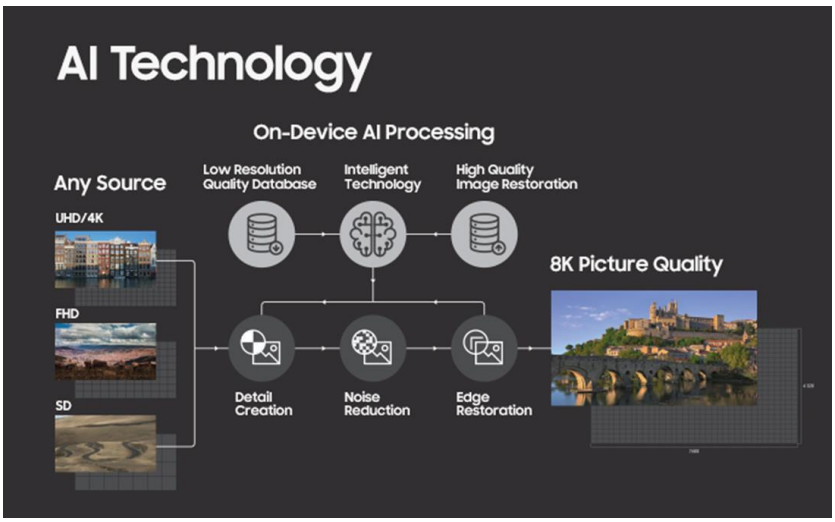
[334] Sur NVIDIA researchers develop AI that removes noise from images with incredible accuracy, juillet 2018.

[335] Voir Super resolution with Generative Adversarial Networks de Boris Kovalenk.

Cela rappelle par la même occasion le scénario de l'excellent film « Sens Unique » avec Kevin Costner, réalisé pendant les années 1980 et qui voit la NSA faire ce genre de reconstitution.

Dans Photorealistic Video Super Resolution, juillet 2018 (15 pages), on en apprend plus sur les méthodes et jeux de données d'apprentissage utilisés par ce genre de GAN. Ils utilisent des vidéos 4K récupérées sur YouTube et réduites en 720p (1280x729 pixels). Ils en extraient 300 000 carrés de 128 pixels de côté dont ils créent des versions réduites de 32x32 pixels qui servent à entraîner le système. Je simplifie ! La startup **Let's Enhance** (2017, Estonie) propose en tout cas déjà d'upscaler des photos à la demande[336].

Le plus étonnant est de retrouver cette technique au CES 2018 chez Samsung qui introduisait cette fonction dans un prototype de TV 8K, capable d'upscaler des contenus SD, HD et UHD en 8K. La méthode utilisée doit être voisine, utilisant un réseau de neurones entraîné à upscaler des morceaux d'images. Mais on attend de voir ce que cela donne à grande échelle. Cela doit bien entendu dépendre de la nature des bases d'entraînement.

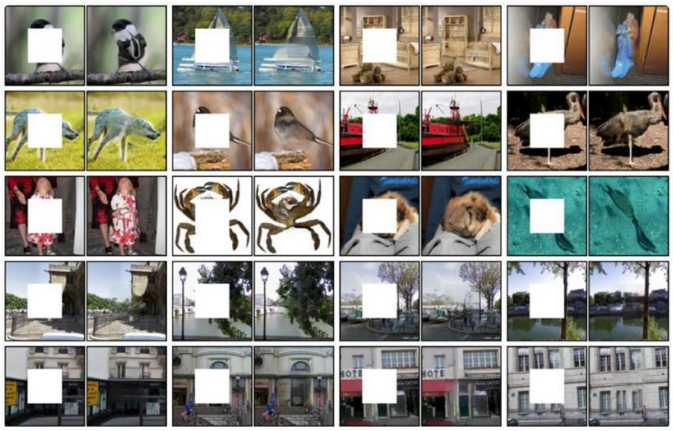


Compléments

Le quatrième grand exercice de style des GANs est de compléter des images incomplètes. C'est une sorte d'exercice de prévision.

Les résultats sont intéressants mais imparfaits, surtout si l'on regarde de très près les images générées, qui par ailleurs sont à basse résolution.

Il ne faut pas oublier que ce sont des modèles probabilistes ! Et lorsque l'effet est de petite taille, comme à droite, vous n'y voyez que du feu ![337]. Cela joue sur les limites de notre cortex visuel qui lui aussi interprète les images générées par la rétine par approximation.



Des zones de formes arbitraires peuvent être également remplies avec des GANs, comme dans Globally and Locally Consistent Image Completion, 2017 (14 pages), illustré *ci-dessous*.

Certains réseaux génératifs sont plus utiles comme WESPE: Weakly Supervised Photo Enhancer for Digital Cameras, d'une équipe de l'ETH Zurich 2017 (10 pages), qui améliore les photos prises par smartphone notamment au niveau contraste et basses lumières, ou celui du stabilisateur vidéo du Pixel 2 évoqué dans Fused Video Stabilization on the Pixel 2 and Pixel 2 XL, 2017 (vidéo).

[336] Elle propose un test d'upscaling, permettant l'upload d'une photo, mais demande ensuite les coordonnées de l'utilisateur. Je ne l'ai donc pas testée. La génération de leads à la petite semaine me fatigue !

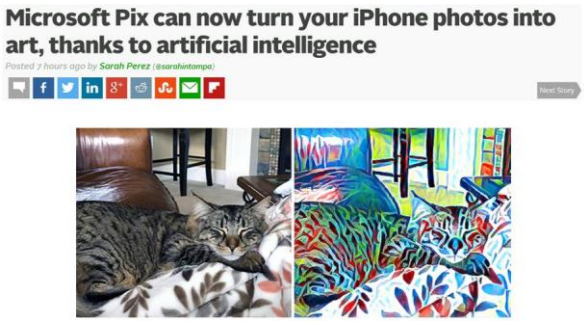
[337] La source de l'image est Context Encoders: Feature Learning by Inpainting, 2016 (12 pages).

D'autres sont capables de modifier les yeux dans une photo, comme dans Eye In-Painting with Exemplar Generative Adversarial Networks, d'une équipe de Facebook, juin 2018 (10 pages).

En 2018, **Nvidia** présentait une solution de ralenti vidéo réalisé à base de GPU et de réseau génératif. C'est impressionnant mais cela doit fonctionner seulement à basse résolution (vidéo). Une autre application consiste à prévoir la suite d'une vidéo, comme vu dans Generating Videos with Scene Dynamics, 2016.

Transfert de style

L'application d'un style à une photo ou une image est une autre application classique des GANs. Elle est souvent montée en épingle comme relevant de la créativité alors qu'il s'agit d'un automatisme déterministe. Ses applications commerciales sont d'ailleurs encore rares.



C'est une application directe des réseaux de neurones convolutionnels capables de détecter des features associés à des autoencodeurs, capables de réencoder ces features à partir d'autres bases. C'est un procédé très mécanique qui n'est pas aussi créatif que l'on pourrait le croire[338] ! L'exemple en bas provient de Li et Wand[339].

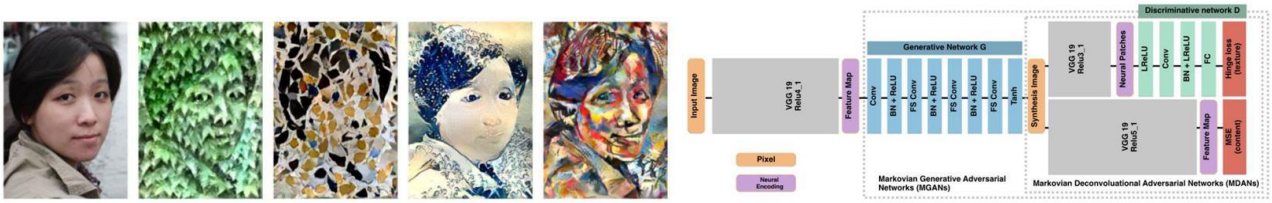
Les cas connus relèvent de l'application du style graphique d'un peintre à une image ou à une photo[340]. C'est plus facile à réaliser avec une image de dessin animé car la plausibilité du résultat est moins remise en question par l'œil humain[341].

---

[338] Voir Can AI make anyone an artists, septembre 2017. On y trouve aussi la vaste plaisanterie pour gogos de pix2code, une AI qui serait capable de créer un programme à partir d'une simple interface utilisateur, la demonstration étant faite avec une interface comportant deux boutons.

[339] Voir StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, 2016-2017.

[340] Voir Painting like Van Gogh with Convolutional Neural Networks, novembre 2016.



Cela devient plus impressionnant avec Image-to-image translation with conditional adversarial networks, 2016 (17 pages) qui génère une image plausible à partir d'un simple schéma (*ci-contre*)[342]. A ceci près que les schémas sont eux-mêmes générés à partir de photos, ce qui facilite très probablement le travail du GAN.

Figure 16: Example results of our method on automatically detected edges→shoes, compared to ground truth.

Autre variante dans le transfert de style, la transformation d'une image d'extérieur prise en hiver en image d'été chez Nvidia dans Unsupervised Image-to-Image Translation Networks, 2018. Là encore, le système a eu besoin d'exploiter une belle base d'entraînement pour générer ce résultat impressionnant.

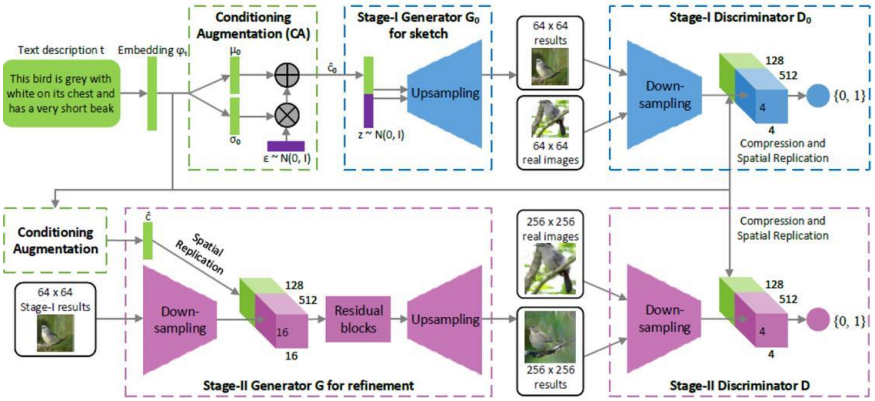
Création ex-nihilo

Les autres formes intéressantes de réseaux génératifs sont ceux qui ont la capacité de générer une image plausible à partir d'un descriptif textuel plus ou moins précis. Il faut évidemment disposer d'une très bonne base d'entraînement pour ce faire avec plein d'images taggées avec de nombreux attributs.

Les images générées sont loin d'être parfaites, mais elles trompent facilement le cerveau dans les exemples fournis par les chercheurs[343].

L'exemple *ci-contre* et *ci-dessous* est issu de Stack-GAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, 2017 (14 pages).

On est en plein réseau de neurones multimodal, associant textes et images !



---

[341] Et puis aussi, cette tranformation d'un cheval en zèbre dans une vidéo (vidéo).

[342] Voir aussi Image-to-Image Demo Interactive Image Translation with pix2pix-tensorflow de Christopher Hesse, février 2017.

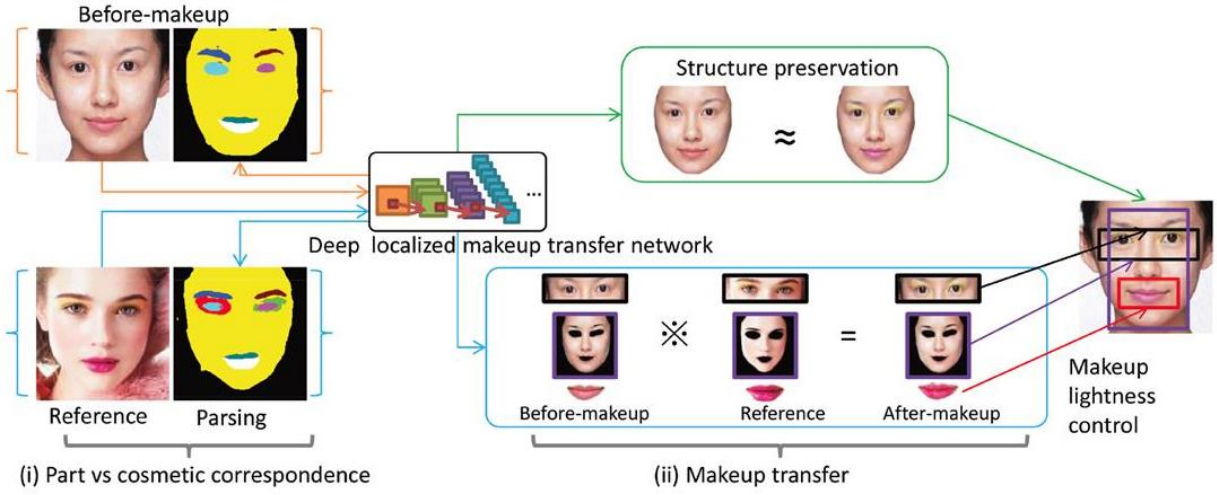
[343] Les publications scientifiques des réseaux génératifs négligent souvent un point clé : la proportion des images générées qui ne sont pas correctes et que le cerveau humain ne reconnaît pas alors que le discriminateur utilisé dans les GAN les a considéré comme des images plausibles.

Visages

Les GANs sont très employés pour modifier ou améliorer des visages. On trouve ce genre de fonctionnalités dans des applications mobiles comme avec l'application mobile **FaceApp** qui peut vous vieillir, vous rajeunir et même changer votre sexe[344].

L'amélioration de selfies fait aussi partie des projets d'**Adobe** qui utilise son IA maison Sensei, exploitant du deep learning pour améliorer de manière semi-automatique les selfies pris avec des smartphones (vidéo) comme pour corriger les perspectives et divers paramètres de prise de vue a postériori.

Le maquillage virtuel passe par une analyse du visage pour le décomposer en parties auxquelles sont appliquées ensuite divers produits de cosmétique[345], qui peuvent ensuite être bien évidemment commandés en ligne. Nombre de startups du secteur proposent maintenant cela. L'une d'entre elles, **Modiface** (2007, Canada) a été acquise par L'Oréal en mars 2018.



Mais une équipe de chercheurs chinois a voulu aller dans le sens inverse, créant une photo d'une femme sans maquillage à partir d'une photo avec maquillage, vu dans Anti-Makeup: Learning A Bi-Level Adversarial Network for Makeup-Invariant Face Verification 2018, (8 pages). Cela marche cependant moins bien que l'ajout de maquillage !

---

[344] Seulement dans la photo !

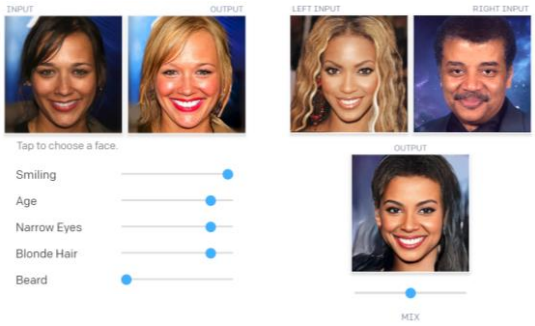
[345] Voir Makeup Like a Superstar: Deep Localized Makeup Transfer Network de Si Liu, Xinyu Ou, Ruihe Qian, WeiWang et Xiaochun Cao, 2016.

La génération d'avatars 3D animés à partir d'une simple photo, utilisant une technique connue de réseau de neurones convolutionnels génératifs. Elle est notamment proposée par la startup américaine **Loom.ai** (2016, $1,35M), créée par des anciens de Dreamworks et LucasFilm (vidéo).

Dans Progressive Growing of GANs for Improved Quality, Stability, and Variation, 2017 (26 pages), une équipe de Nvidia a pu générer des photos de relativement aute résolution (1024 pixels de côté) d'acteurs qui n'existent pas, grâce à un entraînement progressif du générateur du GAN, avec ajout de couches étape par étape pour double la résolution spatiale.

Ce GAN a été entraîné avec le dataset **Celeba**, qui contient des photos de célébrités avec 203 000 photos de 10 177 personnes.

L'un des modèles génératifs les plus récents est **GLOW**[346], créé en juillet 2018 par **OpenAI**. Ce réseau réversible utilise des convolutions de 1x1 pixel. Il permet de générer des images à haute résolution et d'en modifier diverses propriétés comme le sourire, l'âge, la couleur des cheveux et la pilosité. Le modèle permet aussi de mélanger deux visages. Dans le genre glauque, **DeepFake** peut générer des vidéos de porno avec des personnes qui y sont intégrées à l'insu de plein gré[347].



Le buzz a été déclenché avec une démonstration plaquant le visage de l'actrice Gal Gadot sur le corps d'une actrice de porno, habillée pour le besoin du buzz, une fois n'est pas coutume, vu dans AI-Assisted Fake Porn Is Here and We're All Fucked, décembre 2017.

La menace est telle que le Sénateur du Nebraska, **Ben Sasse**, s'en est même ému en octobre 2018 en avertissant de l'impact potentiel de cette technologie sur les débats politiques aux USA, déjà bien mis à mal dans l'ère Trumpienne[348].

Une technique voisine est utilisée à des fins moins répréhensibles, comme dans Everybody Dance Now de Berkeley, 2018 (9 pages) qui fait danser en vidéo une personne qui ne sait pas danser en imitant une autre personne qui danse. Et Google fait quasiment pareil en imitant vos mouvements avec une sacade de photos de personnes qui sont dans la même position, dans **Move Mirror** (vidéo)[349].

Sont donc lancés des projets divers pour détecter et supprimer les DeepFake. C'est le cas du bizarre In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, 2018 (7 pages) qui vise à détecter les bizarreries de clignements d'oeils dans les vidéos pour identifier les fake. Jusqu'au jour où les fausses vidéos imiteront bien le clignement des yeux, ce qui ne devrait pas être trop difficile à faire !

Gfycat (2013, USA, $10M) un site de partage GIFs qui détecte les fake on ne sait pas trop comment puisqu'il s'agit d'une startup et pas d'un laboratoire de recherché qui publie ses travaux.

---

[346] Voir https://blog.openai.com/glow/.

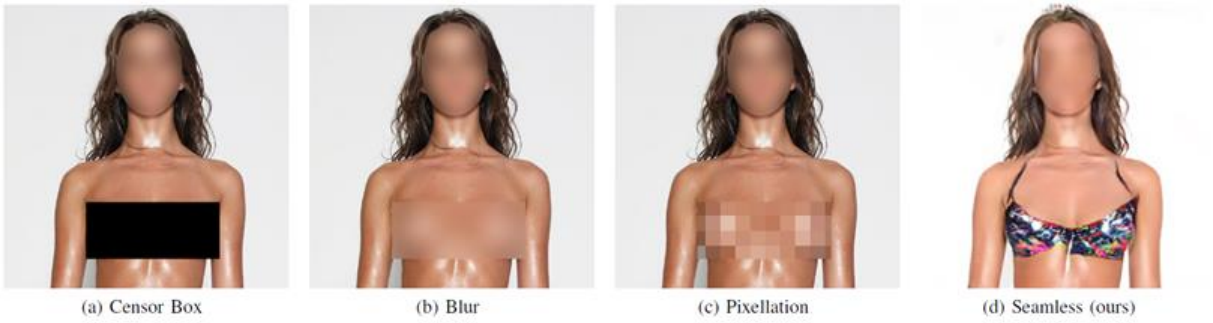
[347] Les vidéos **DerpFake** déclinent la méthode en plaquant diverses personnalités sur des acteurs dans des extraits de films comme pour les James Bond (vidéo), le pire étant, à la fin, avec Rowan Atkinson qui remplace Daniel Craig dans Casino Royale.

[348] Voir This new technology could send American politics into a tailspin, Ben Sasse dans le New York Times, octobre 2018.

[349] Voir Move Mirror: You move and 80,000 images move with you, 2018.

Quant à **Truepic** (2016, USA, $10,5M), ils proposent d'éviter les fake dans les photos ! Comment ? Sans IA. Juste en integrant un watermark dans les photos générées.

Nous avons aussi un projet en cours de **SRI International** et de l'**Université d'Amsterdam** qui est financé par la **DARPA** dans le cadre du programme **MediFor**, pour Media Forensics. Il vise également à détecter des images et vidéos trafiquées[350].



(a) Censor Box (b) Blur (c) Pixellation (d) Seamless (ours)

Fig. 1. Techniques for censoring sensitive regions of an image. *(a)-(c)*: **manual** strategies commonly used for localized censorship. So far, no studies have addressed this problem with an automatic approach. *(d)*: result of our fully-automatic seamless censoring approach using unpaired image-to-image translation.

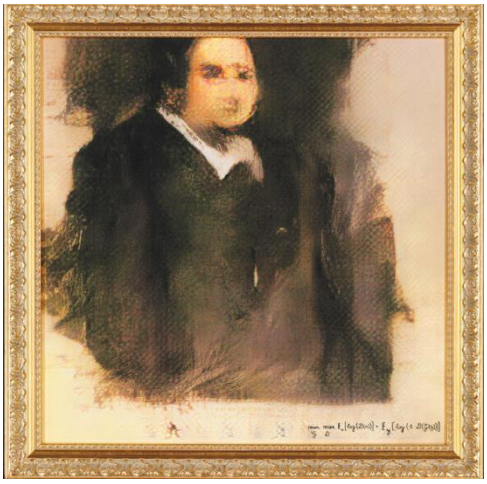
Il y a aussi le contraire, avec ce projet brésilien Seamless Nudity Censorship an Image-to-Image Translation Approach based on Adversarial Training, 2018 (8 pages) qui rajoute automatiquement des bikinis sur les photos de nues, un projet issu d'une Université Catholique financée par le diocèse local.

L'idée est de censurer les images sans que cela se voie trop. Mais ces petits malins démontrent aussi l'inverse, enlevant le bikini (virtuellement) à quelques femmes pour révéler leur tenue d'Eve.

Peinture

La peinture à base de réseaux génératifs donne lieu à des performances plus ou moins convaincantes selon les cas.

Il y a eu par exemple cette génération de tableau artificiel de Rembrandt réalisée avec l'aide de **Microsoft**[351].



En octobre 2018, le premier tableau réalisé à base d'IA générative représentant un personnage fictif, **Edmond de Belamy**, a été vendu aux enchères chez Christie's pour $432K alors qu'il avait été initialement estimé entre $7K et $10K. De quoi faire jaser. L'auteur est le collectif d'artistes français, **Obvious**, qui regroupe Hugo Caselles-Dupré, Pierre Fautrel et Gauthier Vernier. Leur IA à base de GAN a été entraînée avec 15 000 portraits peints entre le 14e et le 20e siècle. Une fois encore, il faut rappeler que l'IA n'a pas réalisé toute seule cette peinture. Ce sont des peintres outillés par l'IA qui l'on créée[352] ! Il a fallut créer du code pour y arriver !

[350] Voir DARPA is funding new tech that can identify manipulated videos and 'deepfakes', de Taylor Hatmaker, avril 2018.

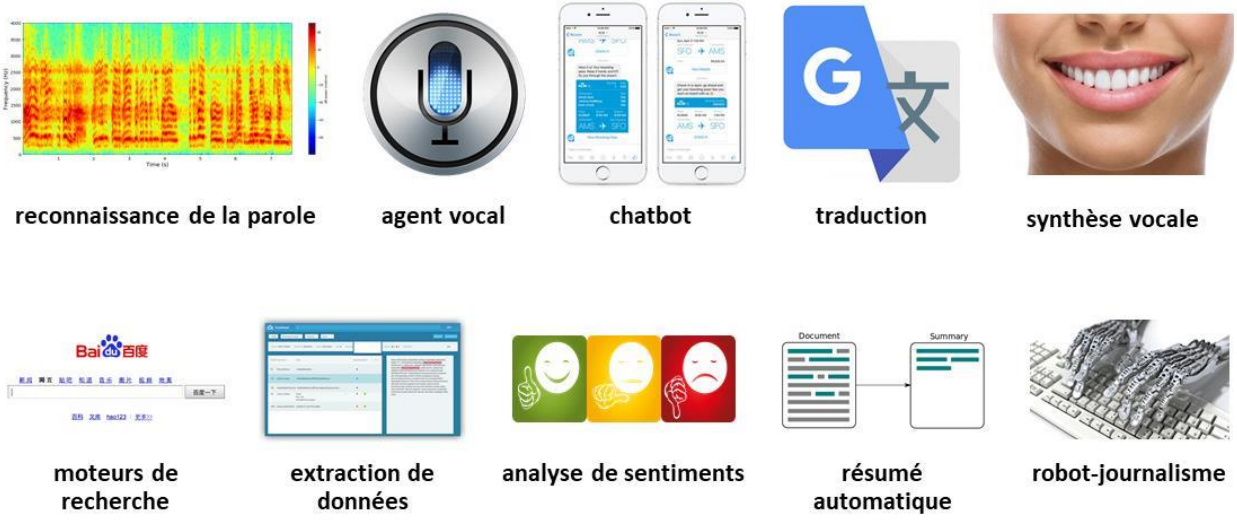
[351] Voir The Next Rembrandt.

[352] Voir Is artificial intelligence set to become art's next medium?, Christie's, octobre 2018. A noter que le collectif Obvious a utilisé une partie du code provenant d'une autre artiste, Robbie Barrat, qui l'avait partagé sous une licence open source (certes, lisible, utilisable et modifiable par tous).

Langage

Le traitement du langage est le second plus grand domaine d'applications de l'IA avec celui de l'image. Il comprend de nombreuses fonctions et notamment la reconnaissance de la parole, les robots conversationnels, la traduction automatique, l'extraction de données, la création de résumés et la génération de textes.

Ces outils couvrent tout le spectre qui va de la compréhension du langage à son interprétation, son exploitation puis à la création de textes ou de paroles. Il comprend aussi les outils et méthodes de représentation des connaissances.



Ce domaine exploite surtout le deep learning et les réseaux récurrents et à mémoire. Ce champ de l'IA est cependant un peu moins mature que celui de l'image. Autant, par exemple, peut-on dire qu'une IA de diagnostic dans l'imagerie médicale équivaut à celle d'un spécialiste, autant un chatbot est encore loin de passer avec succès le test de Turing et d'arriver à se faire passer pour un Humain[353]. Ou tout simplement, à conduire une discussion cohérente de bout en bout pour des demandes élémentaires, même dans l'environnement calme de son logement.

Le deep learning appliqué au langage est aussi probabiliste que celui qui est appliqué aux images. Celui-ci a permis de générer d'énormes progrès dans tous les domaines du traitement du langage, en gros, entre 2012 et 2017[354].

Ces avancées du deep learning sont parfois remises en cause par des chercheurs et entrepreneurs qui trouvent que ces approches probabilistes ont des limites. Ils remettent au gout du jour des méthodes qui réinjectent un peu de symbolisme dans les procédés employés[355]. Ce que l'on retrouve aussi bien chez Google avec ses **Universal Transformers** ou des startups telles que le Français **Golem.ai**, spécialiste des chatbots.

---

[353] Un agent conversationnel est censé avoir passé le test de Turing en 2014. Voir https://en.wikipedia.org/wiki/Eugene_Goostman. Mais en imitant un adolescent de 13 ans dans une discussion assez limitée. Donc, le véritable test de Turing n'est pas encore véritablement passé. Même Ray Kurzweil considère qu'il faudra patienter jusqu'à 2030 pour y arriver.

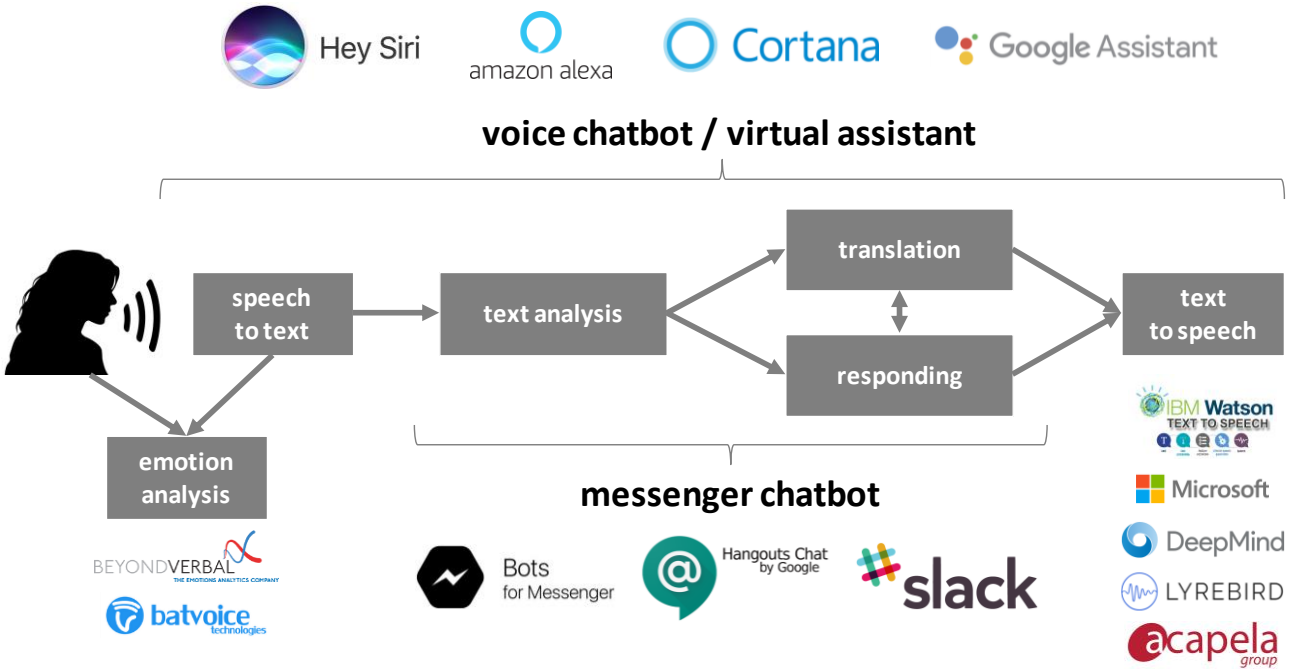
[354] Voir cet excellent historique des avancées dans le traitement du langage : A Review of the Neural History of Natural Language Processing, de Sebastian Ruder, octobre 2018.

[355] Voir par exemple le champ des Memory Networks, 2014, qui gère la mémoire long terme pour les applications de questions/réponses, et les Logic Tensor Networks proposés en 2016 qui combinent les réseaux de neurones avec de la logique symbolique et de la logique floue. What are "Logic Tensor Networks"? de Lucas Bechberger, novembre 2017, Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge, 2016 (12 pages) et la présentation associée Learning and Reasoning in Logic Tensor Networks, mai 2017 (38 slides).

Le champ du traitement du langage et de la représentation des connaissances est très riche et en évolution constante. Il est très difficile à suivre pour les néophytes. Les concepts se renouvèlent sans cesse[356].

Ainsi en va-t-il par exemple des Knowledge Graph Embeddings et des Poincaré Embedding qui sont notamment utilisés par Facebook qui visent à représenter efficacement les liens entre expressions textuelles[357].

Dans le schéma *ci-dessous*, je présente de manière synthétique une partie des éléments qui vont suivrent et qui positionnent le rôle des chatbotx textuels, de la traduction automatique, des assisttants vocaux, des systèmes de speech to text et d'analyse des émotions ainsi que des générateurs de parole synthétique. Tous sont potentiellement reliés les uns aux autres. Ce sont les assistants vocaux qui intègrent le plus grand nombre de briques d'IA et sont les plus difficiles à mettre au point. Il n'est ainsi pas étonnant qu'ils soient surtout maîtrisés par les GAFAMI.



Reconnaissance de la parole

La reconnaissance de la parole est la première étape du dialogue naturel entre un humain et une machine. Elle vise à transformer la voix en texte lisibles par un humain et ensuite, traitable par la machine comme le serait un texte que l'on aurait tapé au clavier.

Elle s'appuyait au départ sur des modèles cachés de Markov et des réseaux probabilistes bayésiens. Ces algorithmes généraient des taux d'erreur assez élevés[358].

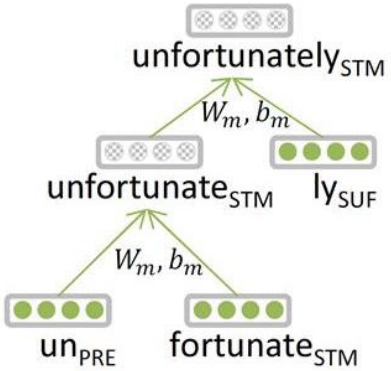
---

[356] Voir Recent Trends in Deep Learning Based Natural Language Processing de Tom Youngy, Devamanyu Hazarikaz, Soujanya Poria et Erik Cambria qui fait un bon état des lieux, février 2018 (24 pages). Les principales techniques de deep learning citées sont le word embeddings et Word2vec (vecteurs de mots qui contiennent les mots apparaissant dans le contexte de mots donnés), les réseaux convolutionnels utilisé pour la modélisation de phrases pouvant notamment servir à la production de résumés, les réseaux de neurones récurrents (RNN, LSTM) qui sont notamment utilisés dans la traduction, les mécanismes de gestion de l'attention qui associent des convnets et des LSTM, les modèles génératifs et les réseaux à mémoire. Pour les plus courageaux, voici Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, de Stanford, 2018 (588 pages).

[357] Voir Knowledge Graph Embeddings: Recent Advances de William Wang, 2018 (21 slides) et Implementing Poincaré Embeddings, de Jayant Jain, décembre 2017.

[358] Pour en savoir plus, voir cet historique de la recherche en reconnaissance de la parole Survey of Technical Progress in Speech Recognition by Machine over Few Years of Research, 2015 (7 pages), qui commence d'ailleurs à dater. Ce sujet intègre de nombreuses branches du savoir issu de plusieurs décennies de recherches dans l'IA.

La complexité de la reconnaissance de la parole provient des variantes dans l'expression qui dépendent du contexte et aussi des ambiguités du langage.Les progrès récents sont dûs à l'adoption du deep learning qui s'est avéré bien plus efficace que les anciennes méthodes[359]. Les systèmes de reconnaissance de la parole utilisent souvent des réseaux convolutionnels pour reconnaître les phonèmes à partie du signal audio qui est analysé sous la forme d'un spectre de fréquences dans le temps[360]. Ils les assemblent ensuite avec des algorithmes qui permettent d'identifier la morphologie des mots, par assemblage de phonèmes.

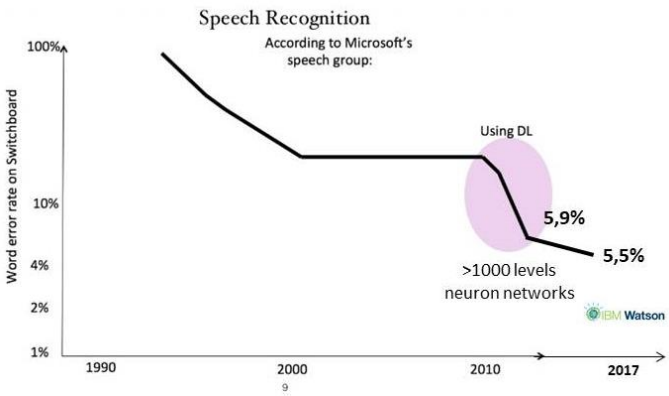


S'en suit, en général avec des réseaux récurrents, l'assemblage de mots dans des phrases ou locutions. Une fois transformés en mots, ceux-ci sont transformés en vecteurs avec autant de bits que de mots dans le dictionnaire, et un seul bit à 1 pour l'indice du mot dans le dictionnaire. Y sont ajoutés ensuite les poids des mots avoisinants dans des phrases connues avec une valeur correspondant à leur probabilité d'apparition autour d'un mot (le mot de référence a donc une probabilité de 1 et les autres, une probabilité inférieure à 1).

Ces vecteurs sont ensuite compressés, (Dense Vector) comme comme via la technique dite Word2Vec, créée par Google en 2013. Cela sert à gagner de la place en ne conservant que les indices et poids des mots pertinents. Ces vecteurs sont utilisés de diverses manières. On peut même réaliser des opérations de logique avec les mots ainsi modélisés (*ci-contre*). Paris moins la France plus l'Italie devient ainsi Rome !

Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

Le reste est une tambouille mathématique de vecteurs qui dépend des nombreuses méthodes utilisées ! Les IA de reconnaissance de la parole manipulent des symboles mathématiques et n'ont qu'une vision purement statistique du langage. Elle n'est pas du tout symbolique. La machine ne comprend pas ce qu'elle interprète. Les progrès de la reconnaissance de la parole se sont accélérés depuis l'utilisation intensive du deep learning avec de nombreuses couches de neurones, jusqu'à 1000 !



Le taux d'erreurs est maintenant inférieur à celui de la compréhension humaine, quc ce soit chez **Microsoft** et **IBM**. Mais il ne s'agit que de taux de reconnaissance de mots il me semble car on peut bien observer au quotidien que les assistants vocaux comprennent à peine deux tiers de nos paroles, comme une grand-mère malentendante.

---

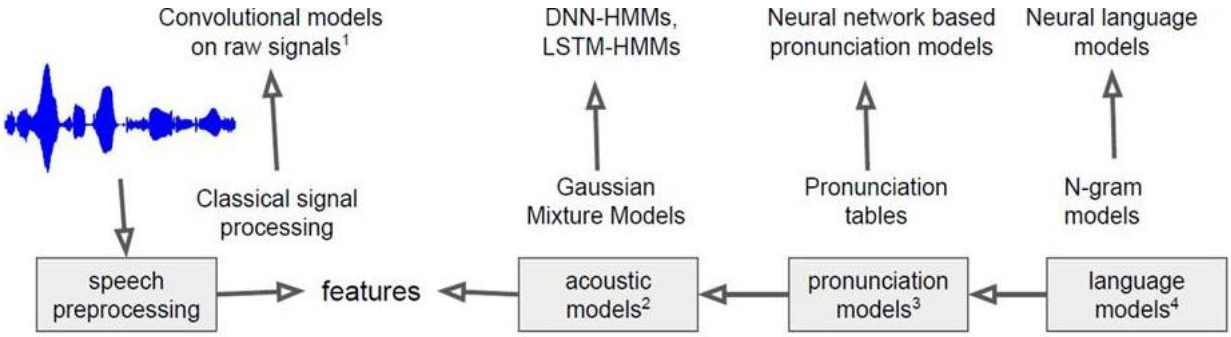
[359] Les premiers progrès en deep learning sont arrivés avec Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, de George Dahl, Dong Yu, Li Deng et Alex Acero, 2010 (13 pages), quis'appuyait encore sur des modèles de Markov. Ces chercheurs de Microsoft Research ont utilisé cette méthode pour faire descendre le taux d'erreurs de la reconnaissance de la parole de 23% à 13% en 2012.

[360] A partir de 2011, on a pu commencer à se passer des transformées de Fourier pour convertir le signal audio en spectre de fréquences, en analysant l'onde audio directement dans des réseaux de neurones. Voir Deep Learning for Speech/Language Processing - machine learning & signal processing perspectives de Li Deng, 2015, slide 123 (200 slides).

Les solutions de reconnaissance vocale à base de deep learning n'ont plus besoin, en théorie, d'être entraînées avec la voix de l'utilisateur. Les modèles de ces réseaux sont créés avec des bases de tests comme pour TIMIT pour l'Américain (qui comprend les mots de 630 locuteurs). Des techniques d'apprentissage par renforcement existent cependant qui affiner la qualité des modèles utilisés.

La reconnaissance de la parole peut être réalisée localement ou sur des serveurs. Avec l'augmentation de la puissance des processeurs embarqués dans les mobiles et même dans certains objets connectés, il est de moins en moins nécessaire de faire un aller et retour avec des serveurs dans le cloud.

Lorsqu'un aller et retour est nécessaire, on voit tout l'intérêt de la 4G et de son débit comme de son faible temps de latence pour les allers et retours avec les serveurs. Et cela sera encore mieux avec la 5G.



1. Jaitly, Navdeep, and Geoffrey Hinton. "Learning a better representation of speech soundwaves using restricted boltzmann machines." *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011.
2. Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal Processing Magazine* 29.6 (2012): 82-97.
3. Rao, Kanishka, et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.
4. Mikolov, Tomas, et al. "Recurrent neural network based language model." *Interspeech*. Vol. 2. 2010.

On est encore loin de la solution parfaite[361], notamment parce que les logiciels manquent d'informations sur le contexte des conversations[362]. Le taux de fiabilité n'est jamais de 100%. Il ne l'est d'ailleurs jamais pour l'Homme également !

Le taux d'erreur de la reconnaissance est toujours plus élevé dans d'autres langues comme le chinois sans compter les langues rares pour lesquelles les bases d'entraînement sont moins grandes que pour les grandes langues parlées dans le monde. D'où l'intérêt de la publication en open source de la solution **Deep Speech 2** de **Baidu** qui fonctionne en anglais et en chinois[363].

Le taux d'erreur est particulièrement élevé dans un environnement bruyant, comme dans la rue, dans un endroit où il y a du monde et même dans sa voiture. Des techniques de captation du son et d'élimination du bruit ambiant permettent de traiter en partie ce problème.

Certaines portent sur l'analyse spectrale et le filtrage de fréquences. D'autres utilisent la captation stéréophonique pour séparer le bruit proche (différentié) du bruit lointain (qui l'est moins). J'avais même vu la start-up israélienne **VocalZoom** (2010, Israël, $12,7M) au CES 2015 qui utilisait un laser pour capter les vibrations des lèvres. Il faut juste trouver où placer le laser, ce qui est plus facile sur des installations fixes que mobiles.

---

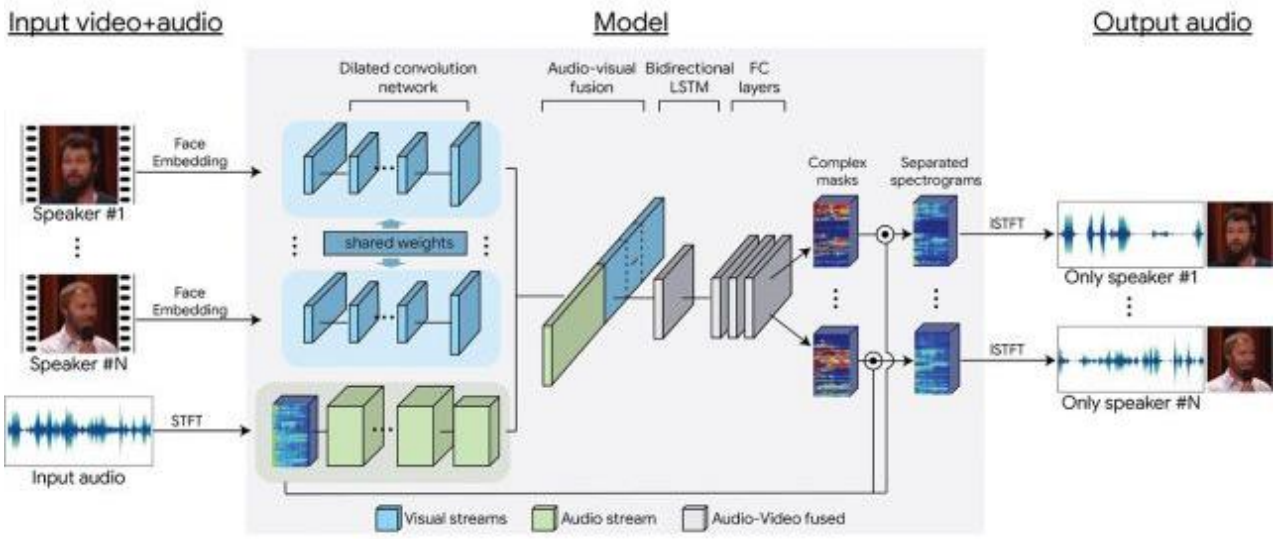
[361] Voir Will computers ever truly understand what we're saying?, janvier 2016.

[362] Voir aussi Why our crazy smart AI still sucks in transcribing speech paru dans Wired en avril 2016.

[363] Voir Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, décembre 2015. Ce système fonctionne avec un réseau de neurones de 11 couches : 3 couches de convolution pour la reconnaissance des phonèmes, sept couches de réseaux de neurones récurrents pour la détection des mots, puis une couche de connexion (« fully connected layer »). En mandarin, il obtient un taux de reconnaissance supérieur à l'homme pour des phrases courtes. Il a été entraîné avec 12 000 heures de conversations. Les versions les plus récentes ont été entraînées avec plus de 100 000 heures de conversations en environnement bruyant.

Une autre manière d'améliorer la reconnaissance de la parole est de faire du « multimodal », à savoir capter plusieurs signaux en même temps comme la voix et une vidéo du locuteur.

C'est ce qu'ont réalisé des chercheurs de Google dans Looking to Listen at the Cocktail Party Speaker-Independent Audio-Visual Model for Speech Separation 2018 (11 pages)[364]. La vidéo associée est très … parlante ! On y voit deux anglophones parler en même temps dans un environnement bruyant. Le système est alors capable d'isoler une à une les deux voix en éliminant l'autre personne et le bruit ambiant. C'est ce que fait notre cerveau lorsque l'on suit plusieurs personnes à la fois dans un diner bruyant !



Le traitement de la parole contient un sous-domaine relativement récent : la détection des émotions dans la parole. C'est l'offre de diverses startups comme **BatVoice** (2015, France) qui se propose ainsi de capter les émotions des clients appelant un call center et d'évaluer l'efficacité des agents qui y répondent et savent traiter le stress des clients.

C'est aussi l'offre d'une autre startup, **BeyondVerbal** (2012, Israël, $10M) qui commercialise de la propriété intellectuelle issue de longues années de recherche dans le domaine. Ils cherchent à détecter des pathologies neurodégénératives avec l'analyse de la parole. Des chercheurs du MIT s'en servent pour détecter des dépressions lors d'entretiens avec des patients[365]. Mais cela reste encore très expérimental.

Il faut se garder de prendre ces solutions à la lettre. Elles sont très approximatives car elles manquent souvent d'éléments de contexte pour bien interpréter les émotions des locuteurs. Les signes extérieurs de ces émotions ne sont pas suffisants pour comprendre les émotions réelles des gens.

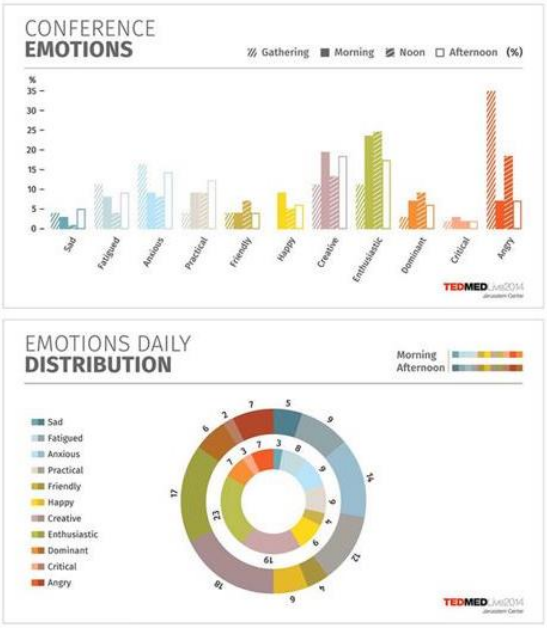
Dans le même ordre d'idée, **Cogito** (2007, USA, $65M) analyse les appels dans les centres d'appels pour fournir un feedback temps réel aux conseillers en ligne. C'est une spin-off du MIT Media Lab qui exploite les sciences comportementales. Ils comparent les caractéristiques des conversations à un historique en analysant la tonalité, le volume, les pauses et la vitesse des discussions. Le système est censé améliorer de 20% la satisfaction des clients.

Dans la même veine, les startups françaises **Natural Talk** (2016, France) et **Cognitive Matchbox** (2016, France) proposent chacune une solution de routage d'appels optimisée aux centres d'appels qui analysent la personnalité et les émotions des clients pour les orienter vers le meilleur agent.

---

[364] Et des explications en langage plus naturel dans Voir Looking to Listen: Audio-Visual Speech Separation, avril 2018.

[365] Voir MIT Develops AI That Can Detect Depression From Your Conversations, février 2018. D'autres utilisent les mouvements du visage pour ce genre de détection, comme le projet de recherche Sensei / Multisence de l'USC (University of Southern California). Voir leur vidéo et leur brochure.



Elles exploitent les APIs d'IBM Watson dédiées au traitement du langage naturel comme Personality Insights, Natural Language Understanding, Tone Analyzer, Document conversion, Twitter Insight et Natural Language Classifier.

AlloMedia (2011, France, 10,8M€) utilise la reconnaissance de la parole pour extraire des informations structurées et semi-structurées des dialogues avec les clients dans les centres d'appels, pour alimenter leurs bases de CRM et améliorer la transformation des leads. C'est ce que propose également **MonkeyLearn** (2014, USA, $1,2M) ainsi que **Dialpad** (2011, USA, $120M).

La reconnaissance de la parole ne permet bien entendu pas de créer une solution complète. Il faut lui ajouter un système qui comprend le sens des questions et qui y répond ! Il doit exploiter une base de connaissance, des arbres de décision et un convertisseur de texte en langage parlé (text to speech). C'est ce que l'on trouve dans les assistants personnels ou chatbots vocaux selon les appellations, que nous verrons plus loin.

Synthèse vocale

A l'autre bout des assistants personnels se trouvent des systèmes de génération de parole synthétique. L'objectif est de rendre les voix artificielles les plus réalistes possibles, ce qui est très difficile à réaliser. On continue à reconnaître les voix synthétiques, même avec les meilleurs outils de synthèse vocale.

Le text-to-speech est une technique complexe, peut-être pas autant que le speech-to-text, mais elle repose aussi sur l'exploitation de réseaux de neurones récurrents, histoire de savoir comment juxtaposer les phonèmes les un aux autres en fonction du contenu à lire.

Google a une excellente solution dans le domaine tout comme **Amazon** avec Polly. Ces solutions sont paramétrables pour spécifier le rythme de la génération, l'intonation, et le style de voix.

Microsoft mettait à jour sa technique de text-to-speech à base de deep learning avec Neural TTS en septembre 2018[366]. La qualité des voix générées est de très bon niveau, quasiment impossible à distinguer de voix naturelles.

Des startups abordent aussi sur ce marché comme la canadienne **Lyrebird** (2017, Canada, $120K) qui propose sa solution aux développeurs sous forme d'API en cloud, exploitant des serveurs à base de GPU Nvidia.

[366] Voir Microsoft's new neural text-to-speech service helps machines speak like people, septembre 2018.

Elle permet de copier la voix d'une personne à partir d'un court échantillon censé être d'une minute et de contrôler l'émotion dans l'intonation[367]. **Acapela Group** (1997, Belgique) propose aussi ses briques logicielles de text to speech qui sont notamment dédiés aux solutions d'accessibilité.

Dans le même ordre d'idée, l'expérience **JFK Unsilenced** de 2018 faisait lire le discours que JFK devait donner après en début d'après-midi je jour de son assassinat à 12h30, au Dallas Trade Mart[368]. L'audio généré est saisissant de réalisme.

La prouesse est cependant techniquement inférieure à celle de Lyrebird avec Barack Obama car elle ne comprend pas la vidéo. Elle est par contre symboliquement beaucoup plus impactante.

Enfin, une équipe de chercheurs de **Baidu** a créé en 2018 une solution de synthèse vocale utilisant la voix d'une personne avec un échantillon très réduit par rapport à l'habitude[369].

Mais pas de démo du résultat en ligne ! L'entreprise de logiciels spécialisée dans le traitement du langage **iFlyTek** (1999, Chine) arrivait de son côté à faire parler Donal Trump en mandarin fin en 2017 (vidéo). C'est un concurrent de l'Américain **Nuance** qui a aussi une offre dans le domaine de l'assistance des juges dans les tribunaux.

Chatbots vocaux

La reconnaisance de la parole est maintenant intégrée dans un nombre croissant de solutions grand public. Le marché est dominé par de grands acteurs américains (**Google** Assistant, **Amazon** Alexa, **Apple** Siri, **Microsoft** Cortana, **Samsung** Bixby qui est probablement originaire de **Viv Labs**[370] sans compter les équivalents chinois comme **Baidu** avec son DuerOS[371].

Leur solutions sont disponibles à la fois dans leurs propres services comme Amazon Echo ou l'iPhone pour Siri mais également disponibles sous forme d'API en cloud exploitables par les développeurs d'applications et de solutions métiers.

Enfin, ils sont intégrés dans des enceintes à commande vocale comme chez Google, Apple avec son HomePod[372], Amazon Echo et plein d'autres copycats originaires d'Asie.

Ces solutions vont d'ailleurs réduire l'intérêt pour certains usages de faire appel à des télécommandes traditionnelles voire même des boutons. Ce sont des plateformes qui proposent un SDK et l'accueil d'applications spécifiques.

---

[367] Voir leurs démonstrations avec les voix de Donald Trump et Barack Obama : https://lyrebird.ai/vocal-avatar. Une vidéo synthétique de Barack Obama a été produite plus tard par l'Université de Washington et le résultat est encore meilleur : Fake Obama created using AI video tool - BBC News. Pendant ce temps là, on n'a pas d'équivalents avec Donald Trump. La résistance au POTUS actuel prend des formes inattendues !

[368] Voir https://rothco.ie/work/jfk-unsilenced/. La performance a reçu le Grand Prix for Creative Data du Cannes Lions International en juin 2018. Voir AI-Driven JFK Unsilenced Triumphs in Creative Data at Cannes de Alexandre Jardine, juin 2018.

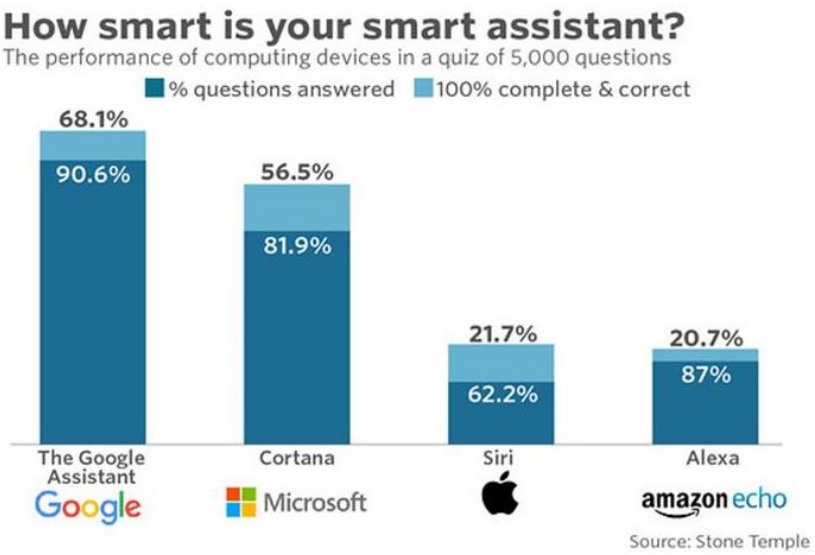
[369] Dans Neural Voice Cloning with a Few Samples, 2018 (17 pages) et les explications dans Neural Voice Cloning with a Few Samples, février 2018

[370] Viv, des créateurs de Siri, est un agent conversationnel capable de répondre à des questions complexes, bien au-delà de ce que peuvent faire Apple Siri et Google. La solution exploite la notion de génération dynamique de programme. Après analyse de la question, un programme complexe est généré en moins d'une seconde qui va la traiter. Viv a été présenté lors de TechCrunch Disrupt à New York (vidéo). Viv Labs (2012, $30M) a été acquis par Samsung pour $215M en 2016.

[371] Voir Baidu Enters the AI Assistant Fray With DuerOS, août 2017.

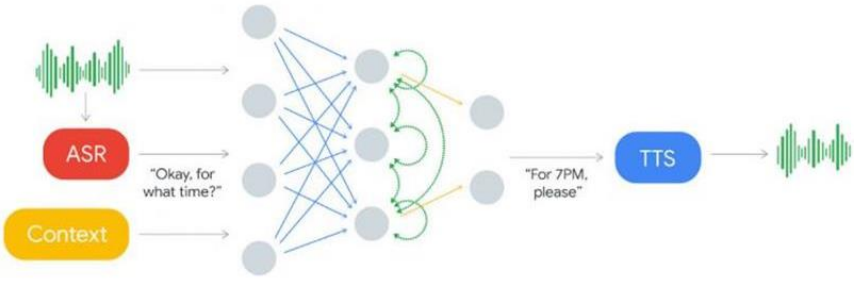
[372] Voir I tried out Apple's new HomePod features. Here's what I learned, mai 2018 qui décrit les fonctionnalités limitées de Siri. L'Apple HomePod est disponible en France et en français depuis juin 2018.

Ces différents assistants se distinguent dans leur capacité à répondre à des questions diverses. Un benchmark de 2017 mettait les assistants de Google et Microsoft devant ceux d'Apple et Amazon (*ci-contre*[373]). Il se trouve que seuls IBM, Microsoft et Google entretiennent une véritable recherche fondamentale sur le sujet, mais ce n'est peut-être pas la seule explication. Le retard technique d'Amazon explique peut-être l'annonce fin août 2016 d'un partenariat entre Microsoft et Amazon.



Ils font en sorte que Cortana puisse dialoguer avec Alexa[374] et réciproquement, rendant ainsi, via la voix, leurs bibliothèques applicatives de services compatibles. On peut donc dire « Alexa demande à Cortana de demander à trucmuch de faire ceci cela ». C'est encore un peu lourd comme forme d'intégration.

Les progrès apparents sont constants avec les agents vocaux. Google a marqué les esprits avec sa démonstration de **Google Duplex** lors de la conférence Google I/O, un assistant vocal qui prend des rendez-vous téléphoniques à votre place[375].



C'est bluffant mais rien ne prouve que cela fonctionne dans une grande diversité de scénarios. Microsoft fait de même, mais seulement en chinois[376], mais l'efficacité est plus difficile à vérifier pour ce qui nous concerne !

La fonction doit faire son apparition dans les smartphones Pixel 3 de Google d'ici fin 2018. Dans les progrès moins spectaculaires, Google Assistant n'a plus besoin d'être relancé par « OK Google » pour une question suivant une autre question[377]. C'est un point clé des agents vocaux : savoir suivre une conversation dans la durée et en préserver les données de contexte.

L'agent vocal peut aussi s'habiller d'un avatar réaliste. C'est le propos d'un projet à venir, **Autodesk Virtual Agent** (AVA) qui est un avatar avec un visage et une voix très réalistes, dont les émotions sont générées par la startup **Soul Machines** (Nouvelle-Zélande, $7,5M), créateurs du Virtual Nervous System (VNS)[378] (vidéo).

---

[373] Voir Alexa and Cortana Will Talk to Each Other Say Amazon and Microsoft, dans Voicebot.ai, août 2017.

[374] Vous pouvez passer un bon moment avec ces deux parodies d'Amazon Alexa par l'émission Saturday Night Live : Amazon Echo – SNL et Amazon Echo Commercial Parody.

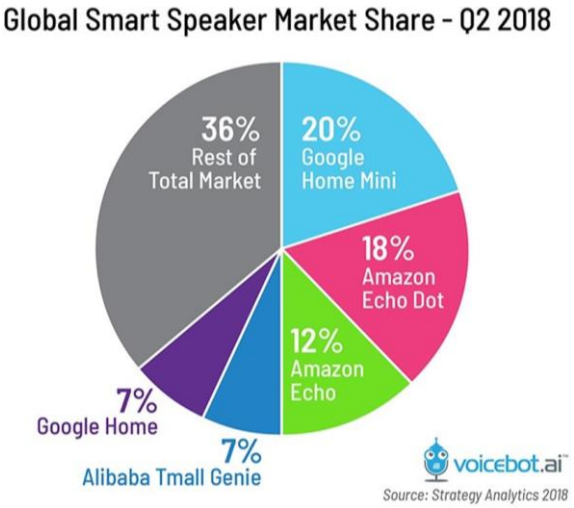
[375] Voir Did Google Duplex AI demonstration just pass the Turing test, mai 2018 et la démonstration en vidéo.

[376] Voir Microsoft's AI Bot Can Make Phone Calls To Humans As Well, mai 2018, mais c'est en Chine et en Chinois !

[377] Voir Google Home now answers follow-up questions without 'OK, Google' wake word, juin 2018.

[378] Voir This Chatbot Is Trying Hard To Look And Feel Like Us, novembre 2017.

Les parts de marché des assistants vocaux ont évolué rapidement en moins de deux ans, surtout côté enceintes connectées. Le marché était dominé par Amazon avec ses Echo, les premiers arrivés aux USA. Mais Google monte en puissance sur le marché des smart speakers depuis début 2018[379], toujours dominé par Amazon côté matériel. Mais en intégrant les agents Amazon Alexa et Google Assistant, la part de marché est plus favorable à Google.



Une nouvelle discipline a fait son apparition : la **VUI** qui est aux interfaces vocales ce que la GUI est aux interfaces graphiques.

La Vocal User Interface d'une application suit le contexte des conversations dans la durée, sait gérer les interactions optimalement, sait reconnaître ses erreurs, etc.

L'américain **Nuance,** qui dépasse $2B de chiffre d'affaire, vend sa solution de reconnaissance de la parole (ASR : automated speech recognition) un peu partout en OEM, notamment dans l'automobile.

Après avoir intégré des technologies de Nuance dans Siri, Apple a fait l'acquisition de la start-up **VocaliQ** (2011, UK) en 2015. **Sensory** (1994, USA) fait avancer l'état de l'art de manière indépendante depuis plus de 20 ans.

Le couteau Suisse **IBM Watson** peut aussi servir à créer sa propre solution pilotée par la parole comme l'a fait l'Américain **Staples** avec son Easy Button qui permet de passer commande de fournitures de bureaux.

Mobvoi (2012, Chine, $252,7M) est une startup de Shanghai proposant un moteur de recherche pilotable par commande vocale. Google Ventures a participé au troisième tour de financement en 2015 avec $60M, lui permettant de mettre un autre pied sur le marché chinois où Google est dominé par Baidu. Le métier principal de cette société est de fabriquer des montres connectées !

Ils ont créé leur Amazon Echo local, le TicHome Mini, de seulement 4,3 cm de haut, 276 grammes et waterproof qui concurrence ceux de Alibaba et Baidu.

Care Angel (2014, USA, $3,4M) propose un asistant vocal qui joue le rôle d'infirmière virtuelle à destination des patients atteints de maladies chroniques.

[379] Voir Google beats Amazon to first place in smart speaker market, juin 2018 qui correspond à la surperfornance de Google sur Q1 et Amazon Takes Top Spot in Smart Speaker Sales in Q2 2018 Says Strategy Analytics, but Google Home Mini Was top Device pour Q2 2018..

Citons enfin quelques agents vocaux qui sont destinés aux objets connectés et commercialisés en OEM par leurs créateurs : Snips, LinTO et Mycroft.

Snips (2013, France, $24,7M) propose un assistant vocal pour objets connectés qui fonctionne en mode autonome sans nécessiter un aller et retour avec un serveur.

Cela lui permet de mieux respecter la vie privée de l'utilisateur, y compris lorsque le service fait appel à des ressources sur Internet. La solution est commercialisée en OEM et à un prix fixe qui ne dépend pas du nombre d'utilisateurs. Ils ont développé aussi un SDK matériel intégrant tout ce qu'il faut pour créer son objet connecté à commande vocale. Tout fonctionne sur l'objet connecté qui peut être un kit Raspberry, à base de processeur Intel ou Arm. Il faut un cœur tournant à un minimum de 1GHz, 521 Mo de mémoire. Le système peut être réveillé avec un mot clé (« Wake word »). Les scénarios utilisent de 500 à 2000 mots pour piloter les objets connectés. L'objectif de Snips est de démocratiser son produit, comme pour permettre la configurer par la voix de son imprimante ou de son routeur Wi-Fi. Ils sont déjà partenaires du français **CareOS**, vu au CES 2018, avec son SDK pour équiper la salle de bain en objets connectés et IA, et intégrés dans le robot ménager multi-fonctions d'une autre startup française, **Keecker**, aussi habituée du CES.

Mycroft (2015, USA, $3M) propose une solution embarquée open source. Ils ont aussi leur « reference design » hardware. Mais une partie de leur logiciel nécessite une connexion au cloud. Ils sont partenaires de Mozilla.

LinTO (France) est une activité de l'ESN Linagora, spécialisée dans l'open source. LinTO est leur agent conversationnel exploitant des ressources en cloud et protégeant la vie privée « by design ». Ils ont notamment la Société Générale comme client. Ils présentaient au CES 2018 un reference design d'enceinte à commande vocale.

Smartly.ai (2012, France, $945K) permet de créer son application vocale facilement. Ils ont déjà la BNP, GRDF et l'Elysée comme client.

Xbrain.io (2012, USA) est une startup établie dans la Silicon Valley ainsi qu'à Paris et Lille qui se spécialise dans les applications de l'IA à l'automobile et la robotique. Sa plateforme xBrain Personal Assistant permet de créer des agents conversationnels, utilisés notamment dans les centres d'appels (sous l'appellation satisfaction.ai) et dans l'automobile (vidéo). Elle s'appuie sur la reconnaissance vocale, sur la gestion de contexte, sur la détection des intentions et la gestion de règles. Elle utilise notamment des GAN (generative adversatial networks) pour la génération de dialogues réalistes. Son créateur, Gregory Renard, planche sur l'IA depuis près de 20 ans. Pour lui, le langage est mille fois plus complexe à gérer que la vision. Il faut être créatif pour utiliser l'IA et assembler des techniques très variées : du machine learning à base de SVM jusqu'à des réseaux de neurones à mémoire LSTM.

Davi (2000, France) crée des agents vocaux « émotionnels » en collaboration avec le laboratoire LIMSI. Ils sont notamment commercialisés auprès de grandes entreprises françaises.

Chatbots textuels

Les robots conversationnels ou "chatbots" sont très en vogue depuis 2015. Des outils permettant d'en créer sont proposés par de nombreuses startups ainsi que dans diverses offres de grandes entreprises du numérique (Facebook, Google, IBM, Facebook, Oracle, …).

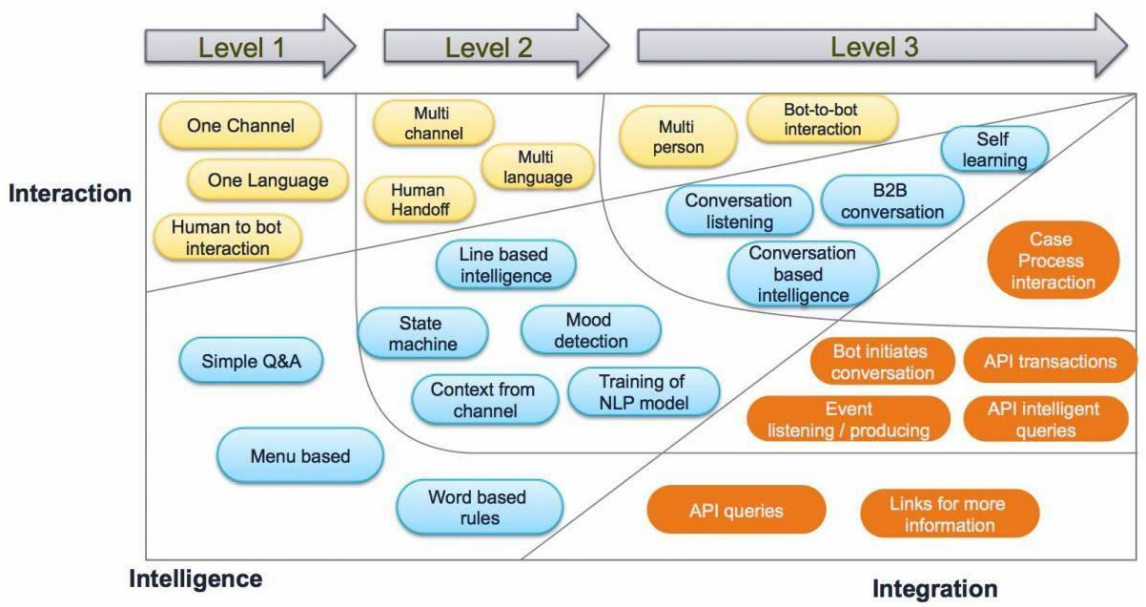
Les chatbots visent à automatiser le service client en ligne dans les sites de e-commerce, services financiers ou autres. Ils sont utilisés également pour des services internes aux entreprises (gestion de notes de frais, demandes de congés, helpdesk IT, conseils juridiques, …). On les utilise à partir d'applications mobiles, de services de messagerie instannée comme Facebook Messenger, ou de sites web divers.

L'objectif ultime est de réussir le fameux test de Turing qui définit une intelligence artificielle comme étant une intelligence indistinctible de celle de l'homme dans de telles discussions par le biais d'échanges textuels. On en est encore loin, même avec les chatbots les plus élaborés.

Ils sont encore très décevants et pas forcément appréciés des utilisateurs. Ces chatbots peuvent avoir une interface vocale comme avec Google Assistant, Apple Siri, Microsoft Cortana et Amazon Alexa.

Il est assez difficile d'évaluer la maîtrise technologique des différentes sociétés de ce secteur. Elles utilisent un patchwork de différentes APIs et outils de deep learning plus ou moins packagés[380].

Certaines redéveloppent leur propre moteur de traitement du langage, ce qui peut paraître curieux en raison de l'abondance de solutions déjà disponibles sur le marché. D'autres se contentent d'un simple moteur de règles, très rudimentaire dans sa portée[381].



Il existe en effet différentes techniques pour créer un chatbot. Elles vont de la gestion basique de questions/réponses à des bots plus sophistiqués capables de trouver de l'information dans des sources variées, de mener des discussions en mémorisant bien leur contexte et de prendre l'initiative, le tout grâce à des techniques avancées de traitement du langage, à des modèles prédictifs et en tenant aussi compte de l'humeur du client. On appelle cela un chatbot de niveau 3[382].

Dans tous les cas de figure, un bon chatbot doit être alimenté par des sources d'information diverses :

- L'accès à des **applicatifs métiers** divers pour interroger des bases de données, faire des réservations, bref, être intégré dans divers systèmes transactionnels.
- L'exploitation d'outils de communication existants avec les clients comme les logs de centres d'appels, les discussions dans les **réseaux sociaux** d'où l'on pourra extraire des dialogues entre personnes réelles pour identifier des réponses à de nouvelles questions.

---

[380] Cet article très intéressant Contextual Chatbots with Tensorflow de mai 2017 décrit comment développer un chatbot avec le SDK de machine learning et deep learning TensorFlow de Google complété par la bibliothèque TFLearn, le tout étant écrit en langage de programmation Python. Tous ces outils sont open source et gratuits.

[381] Voir L'arnaque chatbots durera-t-elle encore longtemps ? de Par Thomas Gouritin, octbre 2017.

[382] Le schéma au-dessus qui décrit les caractéristiques de trois niveaux de chatbots provient de How can Chatbots meet expectations? Introducing the Bot Maturity Model, Léon Smiers, Oracle, avril 2017.

- Des **scénarios d'accueil** et de **questions/réponses** (*exemples ci-dessous*) ce qui peut être très laborieux à saisir si cette connaissance n'est pas déjà formalisée dans l'entreprise ou si elle est difficile à capter.

```
{"intents": [
        {"tag": "greeting",
         "patterns": ["Hi", "How are you", "Is anyone there?", "Hello", "Good day"],
         "responses": ["Hello, thanks for visiting", "Good to see you again", "Hi there, how can I help?"],
         "context_set": ""
        },
        {"tag": "goodbye",
         "patterns": ["Bye", "See you later", "Goodbye"],
         "responses": ["See you later, thanks for visiting", "Have a nice day", "Bye! Come back again soon."]
        },
        {"tag": "thanks",
         "patterns": ["Thanks", "Thank you", "That's helpful"],
         "responses": ["Happy to help!", "Any time!", "My pleasure"]
        },
        {"tag": "hours",
         "patterns": ["What hours are you open?", "What are your hours?", "When are you open?" ],
         "responses": ["We're open every day 9am-9pm", "Our hours are 9am-9pm every day"]
        },
```

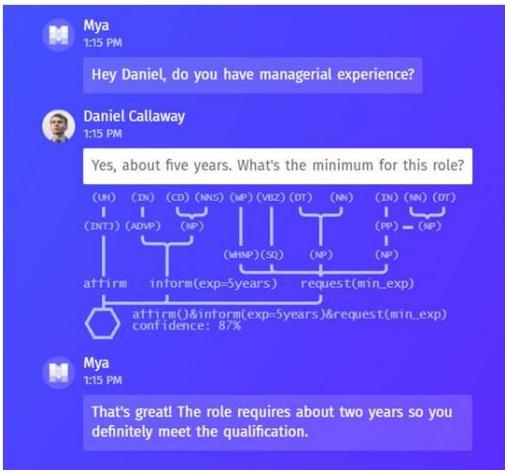

Toutes ces connexions ne se feront pas d'un claquement de doigts ! En général, plus la solution est verticale, moins la startup de chatbot doit disposer de technologie en propre. Ces sociétés se distinguent beaucoup plus par les marchés visés que par leurs choix technologiques ou leurs performances.

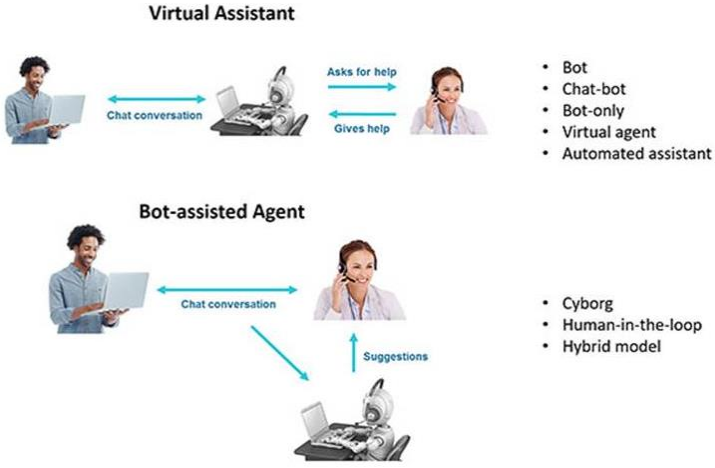
A ce stade de leur développement, les chatbots ne répondent habituellement qu'à des questions très formatées dans un espace sémantique limité au métier de l'entreprise qui le propose. Ils ne savent évidemment pas bien répondre à des questions très ouvertes. Et lorsque la réponse est correcte, il s'agit souvent d'un copier-coller d'une réponse humaine existante dont la grammaire est éventuellement ajustée pour s'adapter au dialogue en cours.

Parfois même, les chatbot génèrent un effet miroir de la bêtise humaine, comme ce fut le cas en 2016 avec le chatbot expérimental de Microsoft Research qui devint rapidement raciste et dû être débranché[383]. En cause, les méthodes d'apprentisssage automatiques exploitant des dialogues avec des utilisateurs. Et c'était avant le 8 novembre 2016 ! Heureusement, les chatbots circonscrits à un domaine métier donné risquent moins de se retrouver dans ce cas-là.

Les chatbots sont de trois types différents du côté des interactions avec leurs utilisateurs :

- Ceux qui fonctionnent de manière entièrement **autonome**. Ce sont des assistants virtuels.
- Ceux qui fonctionnent de manière **semi-autonomes** et sont animés par des opérateurs humains lorsqu'ils ne savent pas bien répondre.
- Ceux qui **aident** des opérateurs humains à répondre aux questions des clients dans les centres d'appels.





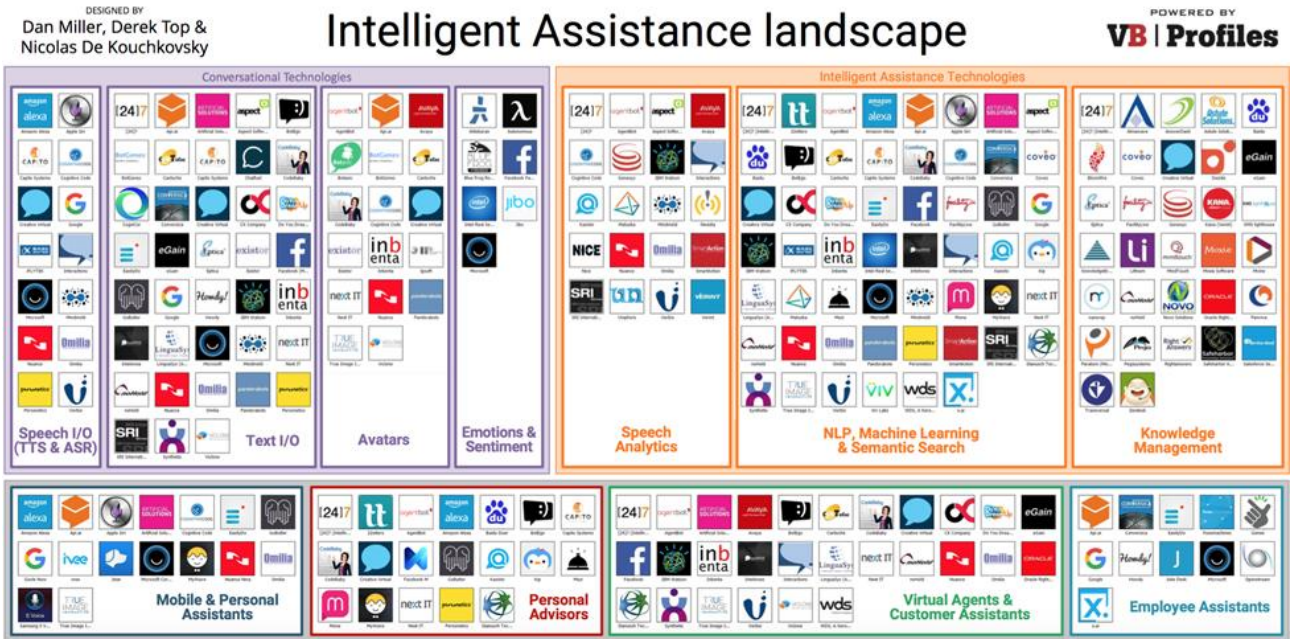
[383] Voir Microsoft is deleting its AI chatbot's incredibly racist tweets, mars 2016.

L'offre peut être segmentée avec des chatbots généralistes, des chatbots spécialisés dans des domaines précis (ecommerce, recrutement, …) et des outils de création de chatbots[384] et des plateformes d'accueil de chatbots comme Facebook Messenger ou Slack.

Les chatbots sont rarement prêts à l'emploi et nécessitent un travail de personnalisation et de mise en place qui est réalisé par l'offreur, par un de ses partenaires services ou par l'entreprise cliente elle-même. On voit d'ailleurs émerger des agences de réalisation de chatbots qui s'appuient sur les outils de création de chatbots du marché.

Le nombre de startups de chatbots créées ces dernières années est impressionnant. Il rappelle la vague des réseaux sociaux après 2004 et celle des services de vidéo en ligne après l'acquisition de YouTube par Google en 2006 !

Leur diversité témoigne d'un marché en ébulition encore immature. En effet, les marchés matures du numérique se distinguent en général par leur sédimentation autour d'un nombre limité d'acteurs. Cela en prend toutefois la tournure avec quelques leaders qui émergent au niveau des plateformes de chatbots.



La platefome de **Facebook Messenger** domine les usages. En effet, les grandes marques et services l'ont choisie parce que Facebook est le réseau social dominant, en tout cas dans les pays développés. Il est suivi de **Slack**, très utilisé pour le travail collaboratif dans les entreprises.

Nous allons faire ici un panorama de quelques-unes des startups de ce secteur en commençant par quelques plateformes de chatbots généralistes :

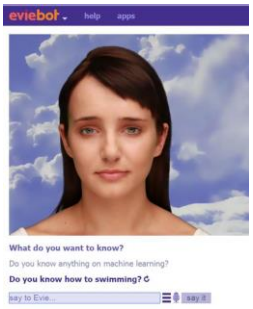
- **Semantic Machines** (2015, USA, $20,9M) est une startup de Boston et Berkeley qui propose des chatbots pouvant être intégrés dans toutes sortes d'usages, b2b et b2c. L'équipe fondatrice comprend des anciens de Siri et Google Now. La solution intègre la reconnaissance et la synthèse de la parole.
- **Talla** (2015, USA, $12,3M) propose une solution de chatbots pour les besoins des entreprises, comme dans le recrutement, le marketing et la gestion de rendez-vous. Elle s'intègre dans les systèmes de messagerie tels que Slack. Elle fait penser au français Julie Desk.

[384] Voir 25 Chatbot Platforms: A Comparative Table par Olga Davydova, mai 2017, qui recense et compare 25 outils de création chatbots.

- De nombreux chatbots spécialisés avec pour commencer, une palanquée de startups dédiée à la création de chatbot pour les sites de vente en ligne : **Msg.ai** (2014, USA, $7,3M) qui est notamment déployée chez Sony, **Niki.ai** (2015, Inde, $2,4M) est une startup indienne qui se focalise dans les services (transports, voyage aérien, santé) et **MessageYes** (2015, USA, $6,5M), une startup de Seattle, qui associe machine learning et opérateurs humains avec deux spinoffs, l'une qui commercialise des disques vinyles (The Edit) et l'autres, des BD (Origin Bound). The Edit aurait vendu $1M de vinyles en huit mois.
- **TARA** (2015, USA, $3M) est une startup de San Francisco qui propose un robot conversationnel de gestion du recrutement de freelances.
- **Do You Dream Up** (2009, France) propose un agent conversationnel multilingues pour les sites web. Il est notamment utilisé par Voyages-SNCF depuis 2011 et a récemment évolué pour être intégré dans une "HelpBox", sorte d'aide en ligne contextuelle interactive. La société a de nombreux clients grands comptes en France, tels qu'EDF dans sa direction juridique[385]. Elle est en train de passer d'une stratégie plutôt services/projets vers une stratégie produit, indispensable pour l'internationalisation de l'activité.
- **Clustaar** (2013, France, $1,7M) a aussi développé une plateforme en cloud de création de chatbots (vidéo).

Passons à quelques solutions de chatbots plus originales :

- **IPSoft** (1998, USA) est une startup assez ancienne de New York qui propose son chatbot Amelia qui est positionné sur le helpdesk IT. Elle exploite un agent de détection d'émotion des utilisateurs provenant d'Affectiva. La startup a fait un pivot et est devenus un chatbot généraliste ciblant la relation client. Mais c'est un marché bien encombré où il est difficile de se différencier !
- **Existor** (1988, UK) a créé Cleverbot qui exploite la webcam des laptops pour interpréter les visages des utilisateurs. Cleverbot utilise la puissance des GPU des ordinateurs et des mobiles. La société propose aussi un avatar visuel pour mener ces conversations. J'ai fait quelques tests et ce n'est pas très probant (*ci-contre*). Et pour cause, les agents conversationnels sont souvent mise en oeuvre dans des univers sémantiques très précis, comme l'offre d'une société donnée. Ils ne permettent pas de naviguer intelligemment dans Wikipedia par exemple !



- **Davi The Humanizers** (2000, France) fait de même avec son chatbot vocal animé par un avatar.
- **Replika** (2014, Russies/USA, $4,54M) développe son chatbot grand public Replika qui joue le rôle d'un ami, conseiller si ce n'est psychothérapeute. Une sorte de "Her", mais pas encore au point. Il sert surtout à choisir des restaurants. La startup explique que sa solution est basée sur une architecture propriétaire de deep learning et qu'elle est dotée d'un fort quotient émotionnel. La startup a été créée à San Francisco par deux russes, dont un spécialiste du traitement du langage.
- **PocketConfident** (USA) est une startup créée à San Francisco par le Français Olivier Malafronte. Elle développe un agent conversationnel spécialisé dans le coaching personnel. En gros, c'est un psy pas cher. C'est un propos voisin de Replika mais avec un angle un peu différent.
- **Gorgias** (2015, France, $3,4M) est aussi positionnée sur l'automatisation du helpdesk IT. Mais son outil aide les conseillers de support à être plus efficaces, sans les remplacer.

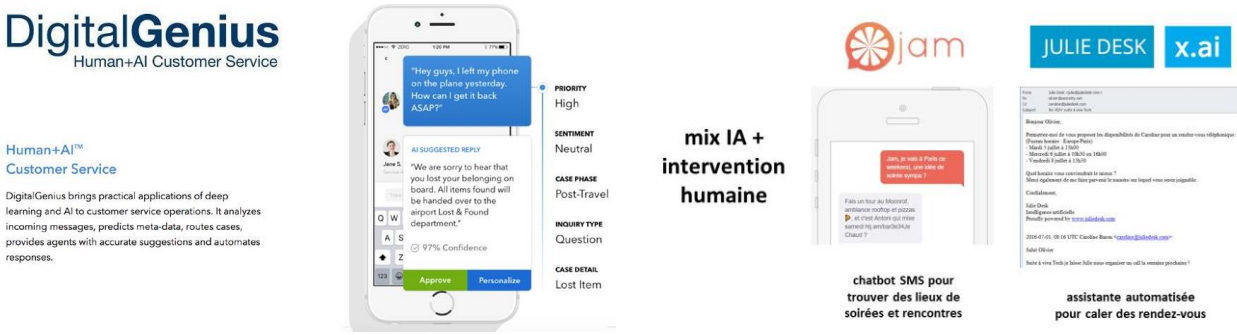
[385] Voir EDF créé un chatbot pour transformer la fonction juridique en entreprise, mai 2018 (vidéo).

- **Publicis** a développé une application de recommandation de maquillage à base de chatbot pour Sephora[386]. Cela relève encore d'une approche de service sur mesure, pas de la création d'un produit.
- **Hubware** (2016, France) utilise une approche intrigante en vendant des assistants conversationnels sur mesure sans technologie en propre, en les assemblant selon les besoins du client.

  Ils apprennent leur métier avec leurs clients, une méthode qui rappelle celle de nombreux cabinets de conseils. A commencer par les sociétés du e-commerce. L'inconvénient de la méthode est que cela rapproche plus la startup d'une société de service que d'une véritable startup à même de générer des économies d'échelle.
- **askR.ai** (2016, France) a développé un chatbot de Business Intelligence servant à interroger des bases de données en langage naturel (vidéo). Reste à voir si cela fonctionne de manière générique pour toutes les bases de données métier (il s'interface avec SAP) et si c'est plus rapide que la manipulation de données dans un outil de business intelligence traditionnel.

Puis aux chatbots associant automatisation et intervention humaine :

- **Curious.ai** (2013, USA, $7,35M) commercialise DigitalGenius qui associe deep learning et intervention humaine pour les chatbots de services clients. Le chatbot qui fonctionne en mode texte sur site web, réseaux sociaux et SMS est entraîné avec des transcriptions d'appels réels au service client.



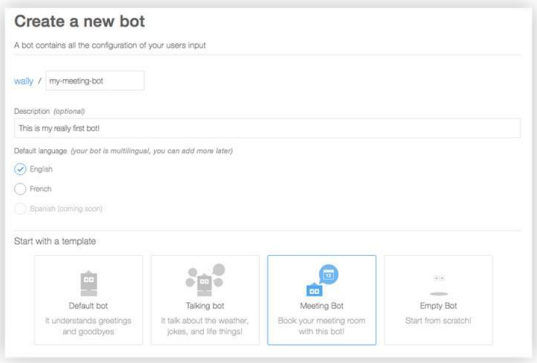
- **Hello Jam** (2012, France, $1,3M) est un agent SMS et dans Facebook Messenger qui envoie une newsletter quotidienne à ses utilisateurs et engage la discussion avec eux. La startup aurait 500 000 utilisateurs. Il ne reste plus qu'à monétiser ! Au départ, c'était un agent SMS pour préparer ses soirées qui était alimenté à la main par des humains et complétés progressivement par divers automatismes.
- **Julie Desk** (2014, France, $3,3M) propose un agent qui répond de son côté aux mails pour l'organisation de rendez-vous. Lui aussi est supervisé par de vraies personnes pour le contrôle qualité. Julie Desk a un concurrent américain, **x.ai** (2014, USA, $44,3M).
- **Kwalys** (France) est une autre plateforme de création de chatbots. Elle mélange arbre de décision et deep learning pour trouver les réponses. Ils sont partenaires d'IBM Watson.
- **Botmind** (2017, France) associe aussi chatbot et humains pour répondre aux clients. Leur solution est notamment déployée chez Axa Banque, BNP Paribas, La Redoute et Prixtel (d'après leur site).

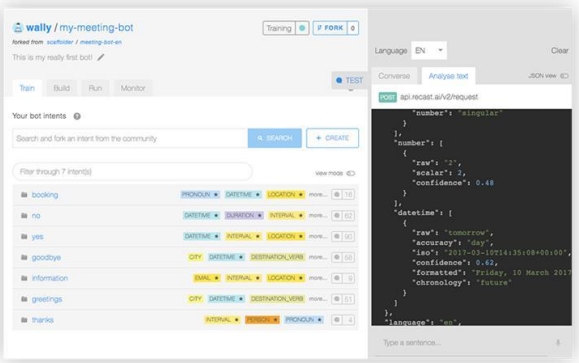
Et enfin, voici quelques outils de création de chatbots exploitables par les entreprises et agences spécialisées de création de chatbots :

[386] Source : keynote de Microsoft AI en septembre 2017, https://myignite.microsoft.com/sessions/56555.

- **recast.ai** (2015, France, $2,25M) est une startup française créée par des anciens de l'école 42 qui propose un outil de création de chatbots et un SDK associé. L'ensemble est très bien packagé et s'utilise en mode cloud (*ci-dessous*). Au cœur de leur solution se trouvent différentes briques internes et externes. L'équipe a pris soin de développer certaines briques de traitement du langage en interne.

  La startup est déjà remarquée aux USA où elle est par exemple très bien identifiée par le fonds d'investissement Andreessen Horowitz dans son Airplaybook[387] comme un acteur clé de la création d'agents conversationnels. La startup a été acquise par SAP en janvier 2018 et sa solution sera intégrée dans SAP Leonardo Machine Learning.





- Le BotMaker de **Viseo** (1999, France, $2M) est un outil de création de chatbots textes et vocaux en open source doté d'une interface graphique interactive capable notamment d'interroger les systèmes d'information des entreprises. Il s'adapte aux grandes plateformes de chatbot comme Facebook Messenger, Slack, Wechat, Amazon Echo et Cortana. Viseo est une société de services spécialisée initialement dans le déploiement d'ERP et qui est devenue un généraliste de la transformation digitale, avec 1200 collaborateurs et 130M€ de CA.
- **Opla.ai** (2015, France), basée près de Clermont Ferrand, propose aussi un outil de création de chatbots. Une partie des logiciels de traitement du langage a été créée par son cofondateur, Mik Bry.
- **Chatfuel** (2016, USA, $120K) est une jeune startup américaine qui permet de créer ses propres chatbots. Sa solution serait déployée chez Forbes, Techcrunch et dans la messagerie instantanée Telegram qui compte plus de 100 millions d'utilisateurs.
- **Pandorabots** (2008, USA) est une startup d'Oakland (Californie) qui propose une plateforme de chatbot en ligne, open source et multi-lingue. 300 000 chatbots avaient été générés au décompte de septembre 2017. Ils sont intégrables dans divers environnements de messagerie instantanés tels que Slack et Whatsapp.
- **Viv Labs** (2012, USA, $30M) est une startup californienne qui propose les outils de création d'assistants vocaux avec des fonctionnalités voisines de celles de SIRI.
- Et bien évidemment, les solutions du domaine issues des GAFAMI, notamment Messenger chez **Facebook** qui permet de développer son propre chatbot, les outils développeurs de Cortana chez **Microsoft,** d'**IBM Watson**, de **Google** et **Oracle**.

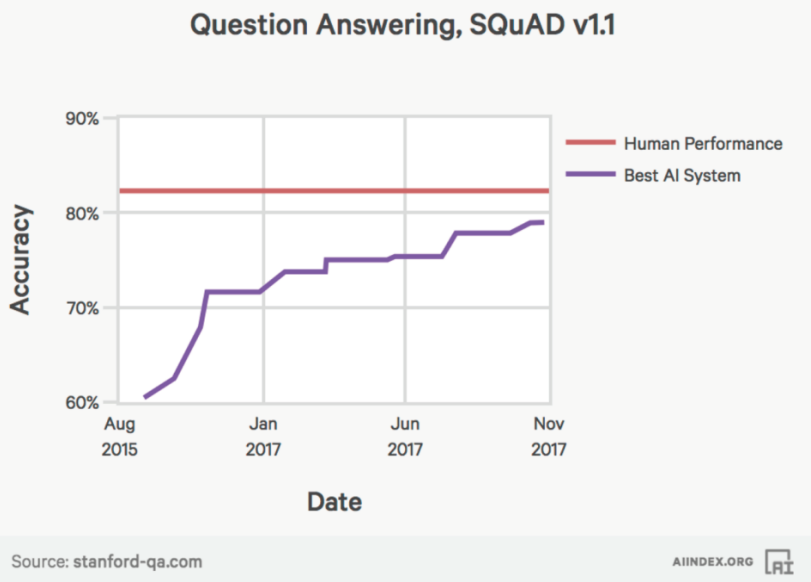
En voici d'autres encore, découverts plus récemment :

[387] Ici : http://aiplaybook.a16z.com/docs/guides/nlp.

- **Botfuel** (2016, France, $4M) propose un SDK de création de chatbots pour développeurs en entreprise.
- **Botnation** (2016, France), basée à Nantes, propose une plateforme en ligne de création de chatbots pour les TPE/PME. Elle a déjà 3500 entreprises utilisatrices. La spécificité ? Des fonctions marketing de tracking.
- **Braina** (2013, Inde) offre la panoplie complète de création de chatbot textuel ou vocal. Cette startup m'a donné du fil à retordre car sur son site et dans la Crunchbase, impossible de savoir d'où elle venait. C'est finalement dans LinkedIn que j'ai découvert que la société était basée en Inde. Eux aussi ont peur de leur ombre !
- **Golem.ai** (2016, France) a développé une technologie de création de chatbot originale qui nécessite moins d'entraînement que les solutions classiques et qui peut fonctionner de manière autonome, le tout sans machine learning ni deep learning. On est donc plus proche d'une IA symbolique. Elle donnerait de meilleurs résultats sur un test publié sur leur site.
- **Omni Ai** (2017, USA, $45M) est un autre original du chatbot qui a développé une IA à base de deep learning autosupervisée, on se demande comment. Et en même temps, ils font aussi de l'analyse de vidéos, suite à l'acquisition de la startup BRS Labs[388].
- **Konverso** (2017, France) a une offre de création de chatbots pour la fonction support des entreprises.
- **LogMeIn** (2003, USA, $30M) propose un chatbot de relation client Bold360 au sein d'une large gamme de logiciels de travail collaboratif pour les entreprises.
- **Wiidii** (2014, France) propose un assistant hybride, avec IA maison et assistants personnels humains. Ils on signé avec un constructeur allemand haut de gamme et ont une application mobile b2c permettant de tester ses capacités.

Il existe même des prix récompensant les chatbots s'approchant le mieux du test de Turing ou le passant entièrement : les **Leobner Prizes**, créés en 1990. S'il a bien été attribué chaque année depuis dans sa première mouture, et notamment au créateur de Cleverbot en 2005 et 2006, il ne l'a pas encore été dans la seconde, celle du passage complet du test du Turing devant deux juges.

A noter que les créateurs de chatbots généralistes peuvent entraîner leur créature avec **Base Squad 2.0** qui est une base de questions/réponses créée en mode crowdsourcing par l'Université de Stanford[389]. SquAD2.0 combine 100 000 questions qui ont des réponses et 50 000 questions qui n'ont pas de réponse. Elles doivent permettre de réaliser des tests de chatbots pour vérifier qu'ils ne répondent pas n'importe quoi aux questions qui n'ont pas de réponse et répondent bien aux autres questions.



[388] Voir la présentation OMNI AI (27 slides).

[389] Voir Know What You Don't Know: Unanswerable Questions for SQuAD, juin 2018 (9 pages).

SquAD a été utilisé par des chercheurs de Microsoft dans **ReasoNet**, une technique associant IA symbolique et connexionniste[390].

Début 2018, Alibaba et Microsoft faisaient réussir le test SquAD par leurs chatbots respectifs, et battre de très peu les capacités humaines[391]. C'est illustré dans le graphe *ci-dessus* qui montre la vitesse de progression dans la création de solutions de réponses au test SQuAD (ici en version 1.1)[392].

Le schéma *ci-dessus* de startups issu de [CES2018] Tsunami of the Voice, janvier 2018, datant du CES 2018, montre que je suis loin de les avoir toutes citées !

Dans les chatbots originaux, il y a le cas d'un chatbot développé pour l'église d'Angleterre à la suite d'un concours organisé par cette dernière en 2017[393] !

Traduction automatique

La traduction automatique s'est longtemps appuyée sur des méthodes statistiques avec énormément de bidouillage manuel. Mise en œuvre à partir des années 1990, elle s'appuyait sur des bases de données, avec d'un côté une base de traduction contenant des expressions et phrases dans deux langues avec des probabilités de correspondance et de l'autre, une base linguistique contenant des textes dans la langue cible corrects d'un point de vue grammatical et du style.

Le deep learning a fait son apparition dans le domaine relativement récemment, vers 2012. Il exploite des réseaux de neurones récurrents (RNN), leur variante à mémoire (LSTM : Long Short Term Memory) et de nombreuses autres déclinaisons[394].

---

[390] Documentée dans ReasoNet: Learning to Stop Reading in Machine Comprehension, 2016 (9 pages) et décrite dans Machine Reading for Question Answering: from symbolic to neural computation de Jianfeng Gao, Rangan Majumder et Bill Dolan, juillet 2018 (31 slides).

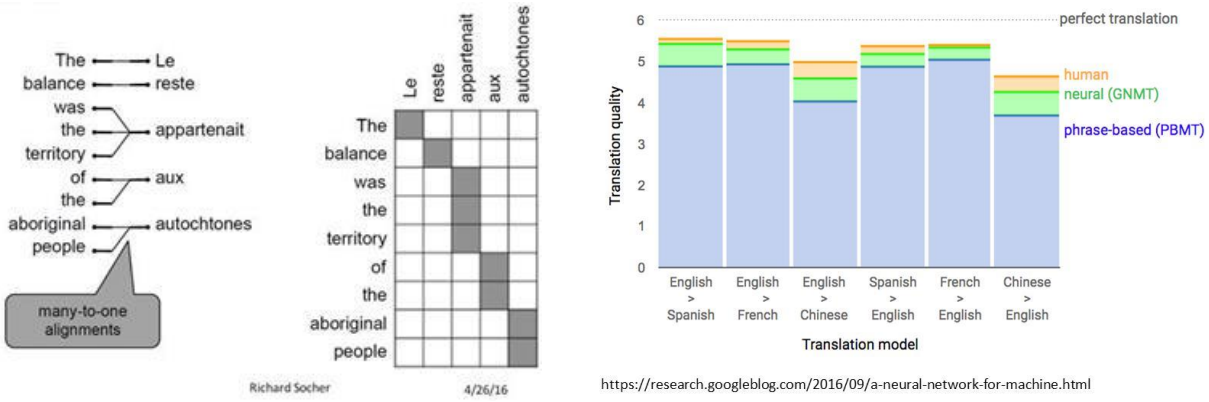
[391] Voir Microsoft, Alibaba AI programs beat humans in a Stanford reading test, janvier 2018. Il semble que les chatbots qui ont réussi cette prouesse aient été réalisés sur mesure et ne sont pas les chatbots commerciaux d'Alibaba et Microsoft. On est encore dans un scénario d'AI étroite !

[392] Issu de Artificial Index Report 2017 (101 pages).

[393] Voir Chatbot for churchgoers among winners of first Church of England Digital Labs competition, octobre 2017.

Ces nouvelles techniques permettent de mieux gérer la traduction en respectant le contexte des mots, expressions et phrases. Cela a permis de passer du presque mot à mot à phrase à phrase, en tenant compte du contexte. Ce champ d'application s'appelle en américain le Neural MT pour **Neural Machine Translation**. Ce courant est devenu dominant en 2016.

Contrairement à la reconnaissance d'images où l'IA a dépassé les capacités humaines, la traduction à base d'IA n'y est pas encore, tout comme les approximatifs chatbots. La traduction à base d'IA est encore imparfaite mais elle réalise des progrès constants, les langues asiatiques étant toujours plus difficiles à gérer car elles sont plus imagées que les langues européennes. D'où la performance remarquée de Rick Rashid, à l'époque patron de Microsoft Research, lorsqu'il démontra en Chine une solution de traduction orale de l'anglais au chinois en 2012[395].



Les systèmes de traduction les plus sophistiqués sont ceux qui font du speech-to-speech, à savoir qu'ils interprètent la voix et non du texte et le transforment en voix dans la langue cible. Ils alignent donc au minimum trois agents : speech-to-text, traduction puis text-to-speech. Ce dernier agent peut d'ailleurs lui aussi s'appuyer sur du deep learning pour générer une voix aussi réaliste que possible[396].

L'un des leaders mondiaux indépendants de la traduction est **Systran**. Cette société américaine créée en 1968 avait démarré en traduisant du russe en anglais pendant la guerre froide. Elle est devenue française en 1986 puis acquise par le coréen **CSLi** en 2014. Elle faisait moins de 10M€ de chiffre d'affaire en 2009. En 2018, l'éditeur lançait Pure Neural Server, une nouvelle génération de solution de traduction automatique destinée aux entreprises construite autour de du framework open source maison de deep learning **OpenNMT** lancé en 2016.

Ils ont aussi une offre spécialisée dans la traduction de documents financiers pour la mise en conformité avec la règlementation européenne MiFID II (Markets in Financial Instruments Directive II).

Google et **Microsoft** proposent chacun de leur côté un système de traduction automatique avec l'application mobile Google Translate d'un côté et Cortana de l'autre.

Google Translate a la particularité de traduire le texte photographié dans des images. Il est également disponible sous forme d'un service Internet capable de gérer des dizaines de langues. Google Translate a fait d'énormes progrès fin 2016 avec l'intégration de son système GNMT (Google Neural Machine Translation) qui exploite massivement du deep learning. Puis c'était au tour de Transformer en 2017 et d'Universal Transformer en 2018. Leur particularité est d'accélérer

---

[394] Voir la conférence Traduction et traitement de la langue naturelle d'Huggo Schwenk dans la chaire de Yann LeCun au Collège de France en avril 2016.

[395] Visualisable ici : https://www.youtube.com/watch?v=Nu-nlQqFCKg.

[396] La méthode est documentée dans la présentation Deep Learning in Speech Synthesis de Heiga Zen de Google, 2013 (48 slides).

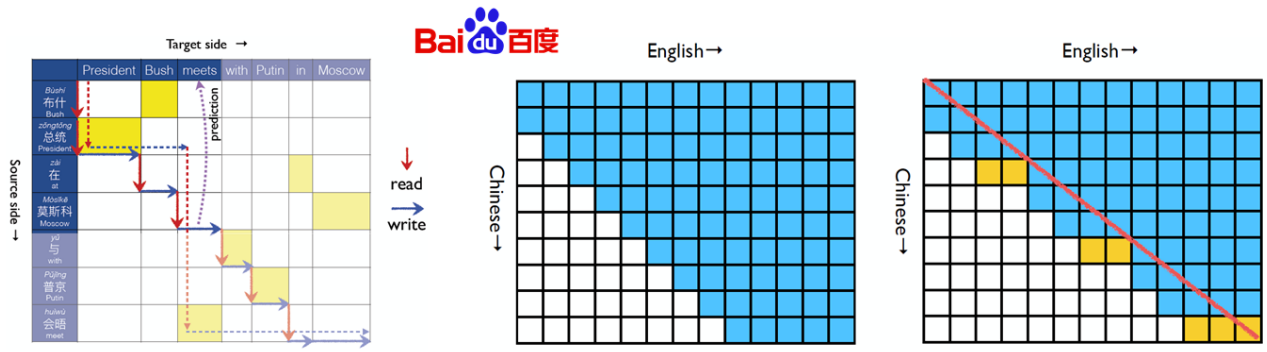
l'entraînement du réseau de neurones en parallélisant les tâches pour chaque mot dans leur contexte, et d'améliorer le résultat de la traduction. Il passe plus de temps à traiter le cas des mots ambigus dont le sens dépend le plus du contexte de la phrase[397].

Cortana se focalise de son côté sur le speech-to-speech. Microsoft teste depuis le printemps 2018 une version de son système de traduction fonctionnant en local sur smartphones, notamment sur les Huawei et leur processeur Kirin 970[398].

Pour sa part, **Facebook** a développé Fairseq, un SDK open source de traduction[399] qui s'appuie sur un réseau convolutif au lieu d'employer des réseaux récurrents comme les systèmes de deep learning habituels (même si nombre de laboratoires et startups s'en éloignent de plus en plus).

Les équipes de recherche du FAIR (Facebook AI Research, pilotées par Yann LeCun) ont notamment créé Rosetta, une technique qui améliore la qualité de la traduction en exploitant des images contenant du texte[400]. C'est une forme de deep learning multimodal comme celui qui est utilisé pour la reconnaissance de la parole exploitant la voix et la vidéo du locuteur.

Baidu est aussi un acteur de ce créneau avec des avancées constantes de sa recherche dans la traduction simultanée. Récemment, leurs laboratoires ont démontré un système de traduction qui réduit le temps de latence de la traduction grâce à un modèle prédictif des mots qui vont être prononcés par le locuteur à traduire. Avec leur STACL, la traduction n'a plus qu'une latence de cinq mots avec un taux d'erreur acceptable. Le système gagne en fait environ 2 à 3 mots de latence ce qui n'est pas grand-chose mais peut servir dans la traduction simultanée dans les instances internationales et dans des contextes politiques sensibles[401]. C'est un projet de recherche, donc pas encore opérationnel.



Quittons les grands acteurs du numérique et passons à quelques startups du domaine de la traduction. Un bon nombre d'entre elles s'appuient sur du crowdsourcing pour améliorer leurs bases ou pour fournir un service de traduction humain assisté par des outils[402].

Lilt (2015, USA, $3M) propose un système de traduction destiné aux traducteurs professionnels.

---

[397] Les détails sont dans Moving Beyond Translation with the Universal Transformer, août 2018, mais je ne vous cache pas que je n'ai pas tout compris !

[398] Voir Microsoft Translator gets offline AI translations de Frederic Lardinois, avril 2018.

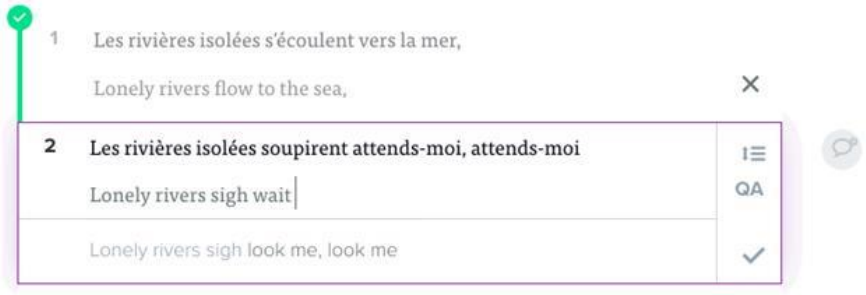
[399] Voir Announcing Fairseq, mai 2017.

[400] Voir Rosetta: Understanding text in images and videos with machine learning, septembre 2018.

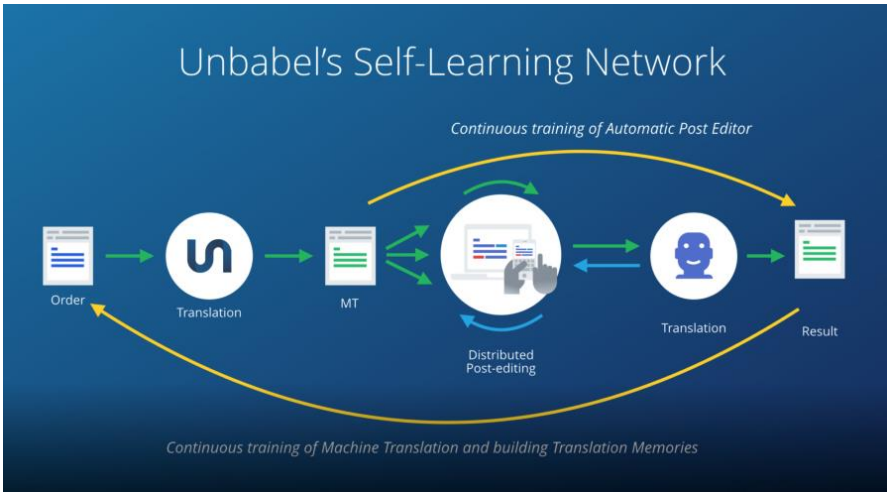
[401] Voir STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency, octobre 2018 (10 pages).

[402] Ces startups extraites de 7 Machine Language Translator Startups, mars 2018.

L'outil de saisie prédictif propose de suggestions au traducteur pendant son travail. Comme le traducteur ajuste le texte qui est proposé, cela permet en retour d'alimenter le système, ce qui constitue une forme d'apprentissage par renforcement.



Unbabel (2013, Portugal, $31M) associe le deep learning et l'intervention humaine, qui améliore de manière continue le système. La startup indique utiliser une communauté de 50 000 traducteurs indépendants couvrant 28 langues. Leur outil s'intègre dans Salesforce et Zendesk et dans d'autres logiciels métiers via leur API.



L'intervention humaine peut être vue à la fois comme un moyen d'améliorer la qualité de la prestation et une limitation intrinsèque de leur IA. On est en droit de douter.

SDL (1992, UK) est un acteur établi de la traduction avec un CA de $490M qui propose, entre autres choses, une solution serveur de traduction pour les entreprises, à base de deep learning. La société fournit aussi des services de traduction traditionnels, ou assistés par ses propres outils.

DeepL (2009, Allemagne) s'appuie sur un supercalculateur pour traduire un million de mots par seconde dans 7 langues avec des résultats qui dépasseraient ceux de Google en qualité. Au prinemps 2018, DeepL s'associait avec **Quantmetry** (2010, France) pour traduire un ouvrage scientifique de 800 pages, de l'anglais vers le français. La traduction était bouclée en deux mois et demi, corrections humaines comprises[403].

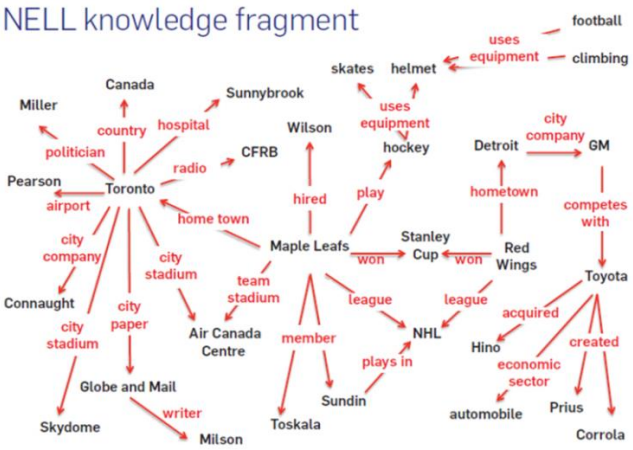
Extraction de données

Le traitement du langage a d'autres applications diverses consistant à exploiter les données qualitatives ou quantitatives extraites de textes.

Elles permettent d'identifier des personnes, des sociétés, produits, lieux, prix ou dates dans des textes et notamment sur Internet et dans les réseaux sociaux. Elles classifient ces informations selon divers critères comme les sujets de discussion ou la tonalité, que l'on appelle l'analyse de sentiments. Les techniques d'IA sont de plus en plus courantes dans ces applications.

[403] Il s'agissait de l'ouvrage de référence « Deep Learning" de Ian Goodfellow, Yoshua Bengio et Aaron Courville. Mais deux mois et demi, divisé par deux pour 400 pages, c'est encore trop pour traduire dans le sens inverse le Rapport du CES en anglais !

Dans ce domaine, l'une des initiatives les plus marquantes est le projet **NELL** (Never Ending Language Learning) est un service d'identification des liens sémantiques entre expressions textuelles lancé par Carnegie Mellon en 2010. Il scanne automatiquement des millions de sources de textes, notamment sur Internet, pour créer sa base. Le projet a été cofinancé par la DARPA et Google. Il génère une sorte de base de contenus sémantiques reliés entre eux. Comme son nom l'indique, le projet n'est jamais terminé et continue sans cesse de s'enrichir.



Les datasets (en anglais) sont librement téléchargeables pour être exploités dans des applications de traitement du langage[404]. Ils comprennent plus de 2,4 millions d'entités.

L'exploitation de réseaux de neurones à mémoire (RNN et LSTM) et leurs variantes et combinaisons permet aussi d'améliorer la capacité à détecter les sentiments dans des textes, de manière plus fine qu'avec une simple analyse syntaxique classique[405]. Ces techniques servent notamment à détecter les discours de haine sur Internet[406].

De nombreuses startups et entreprises opèrent dans ce secteur.

IBM propose de nombreux outils de traitement du langage dans sa boîte à outils Watson, dont DeepQA, qui permet non seulement de créer des agents conversationnels, mais aussi d'extraire des règles issues de textes, comme des documents scientifiques. Cela permet d'alimenter en retour des moteurs de règles pour du raisonnement automatique.

Cinnamon (2012, Japon, $10M) fait aussi de l'extraction d'informations de documents qui permet à leur "Lapis Engine" de faire des recommandations ciblées. Ils ont aussi un "Flax Scanner" qui scanne des documents et extrait les informations non structurées qu'ils contiennent.

Heuritech (2013, France, $1,1M) propose sa solution logicielle Hakken d'analyse sémantique, de tagging et classement automatiques de textes, images et vidéos sous forme d'APIs.

Ils proposent aussi HeuritechDIP qui permet d'améliorer sa connaissance des clients et d'anticiper leurs besoins, évidemment, surtout dans les applications de commerce en ligne et le fashion. Le tout exploite force marchine et deep learning. La startup s'appuie sur les travaux de recherche du CNRS LIP6 et de l'ISIR de l'UPMC (Paris VI). Les deux fondateurs ont un doctorat en IA. Ce n'est donc pas de l'IA washing a priori !

Proxem (2007, France, 1M€) propose Proxem Studio, une solution de traitement automatique du langage permettant de filtrer, analyser, tagger et classifier automatiquement de gros volumes de données textuels, comme dans les commentaires d'utilisateurs dans les réseaux sociaux ou sites de

[404] Voir Aiming to Learn as We Do, a Machine Teaches Itself de Steve Lhor, 2010, Never-Ending Learning, mai 2018.

[405] Voir OpenAI sets benchmark for sentiment analysis using an efficient mLSTM, avril 2017.

[406] Voir Internet researchers harness the power of algorithms to find hate speech, octobre 2017.

e-commerce. Le tout s'appuie sur des techniques de machine learning et de deep learning. L'outil permet notamment d'explorer les données analysées de manière visuelle pour identifier des patterns et signaux faibles. Elle s'est fait remarquer en 2016 en étant utilisée par l'équipe de campagne d'Emmanuel Macron pour analyser le contenu des enquêtes terrain et faire ressortir les thématiques clés.

Klaxoon (2014, France, $55,6M) analyse les textes des idées produits par les groupes de travail dans leur outil de travail collaboratif Brainstorm, afin de proposer des actions pertinentes associées, comme la création de rendez-vous automatique. Klaxoon n'intègre au passage avec Microsoft Teams, la solution de travail collaboratif de l'éditeur américain. L'annonce en était faite en session plénière lors de l'événement Microsoft Experience à Paris le 7 novembre 2018.

Keluro (2014, France), créée par des anciens de l'ENS, se focalise de son côté sur l'exploitation des emails d'entreprises pour en tirer des informations exploitables et structurer les conversations. Ils exploitent des techniques de machine learning pour la classification des informations. La solution est disponible dans Google Play et sous forme d'extension de Microsoft Office.

Lexistems (France) propose une API, X-Act.ai, qui permet l'interrogation et le traitement de données et documents par le sens. Elle permet de consolider facilement des informations et analyses à partir de sources de connaissances multiples et hétérogènes. Je les ai croisés à Nantes au Web2day 2018 ! Cela serait sympathique que leur solution puisse permettre de produire automatiquement le Rapport du CES ou ce genre d'ebook !

Qemotion (2015, France, 614K€) extrait et analyse les réactions textuelles d'une audience pour les convertir en analyses des sentiments.

re:infer (2015, UK, $3,5M) analyse tous les flux textuels et de communication de l'entreprise pour identifier les actions à gérer dans la relation client. Ils sont notamment fournisseur de la solution de RPA **Blue Prism**.

Tessian (2013, UK, $16,8M) est positionnée un créneau voisin de celui de Keluro. Ils analysent les emails sortant des entreprises pour bloquer l'envoi d'informations qui pourraient être sensibles (confidentielles, mauvais destinataire, …).

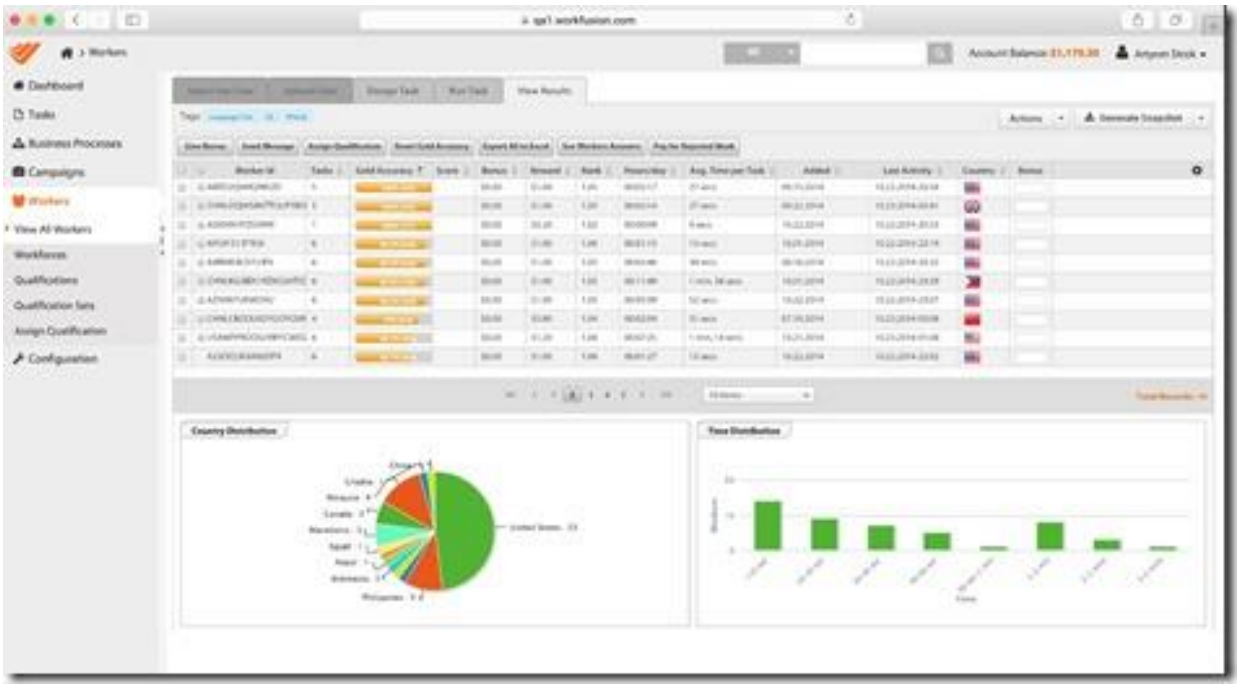
SkipFlag (2014, USA) exploite avec Rover les données circulant dans les entreprises notamment dans les outils collaboratifs comme Slack pour créer des bases de connaisances capables de répondre aux questions clés. Encore une application exploitant l'analyse du langage.

Wrapidity (2015, UK) a développé une technologie pour automatiser l'extraction de données à partir de contenus Web non-structurés. Elle a été acquise par la société de datal analytics **Meltwater** (2014, USA, $60M) début 2017.

Idibon (2012, USA, $6,9M) analyse les textes structurés, notamment issus des réseaux sociaux, pour les classifier automatiquement et réaliser des analyses statistiques dessus.

WorkFusion (2010, USA, $121M) propose l'automatisation de l'exploitation de gros volumes de données non structurées. Il donne l'impression de récupérer les documents comme le fait IBM Watson dans ses outils d'ingestion. Il est par exemple capable de récupérer les résultats financiers de nombreuses entreprises et d'en présenter une synthèse. La méthode relève de la force brute au lieu d'exploiter la chimère du *web sémantique* qui n'a pas vraiment vu le jour. Comme le web sémantique demandait un encodage spécifique et structuré des données, peu de sites l'ont adopté et

l'extraction de données reste empirique. Le traitement même de ces données pour les interroger n'a pas l'air de faire partie de leur arsenal.



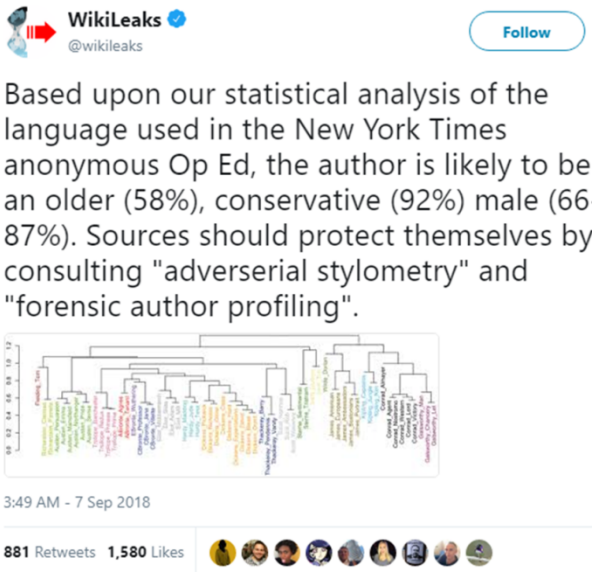
DefinedCrowd (2015, USA, $12,5M), capte les données vocales ou textuelles et les exploite, notamment pour de l'analyse de sentiments. On n'échappe pas à sa **vidéo** avec son ukulélé de circonstance.

Weotta (2011, USA) met en œuvre ce genre de technique dans son application WeottaGo, une application de recommandation mobile.

Il existe évidemment divers outils de développement spécialisés dans le traitement du langage. On peut citer notamment la bibliothèque open source **Gensim** écrite en Python qui sert notamment à analyser des textes, à identifier des sujets traités ou des sentiments et peut-être notamment exploitée dans des applications de commerce en ligne.

L'extraction de données d'un texte devrait aussi permettre en théorie d'identifier leur auteur. C'est ce qu'on tenté plusieurs organisations pour identifier l'auteur de la fameuse tribune libre publiée en septembre 2018 dans le New York Times, d'un membre de l'administration Trump critiquant sévèrement le Président. **Wikileaks** a déduit de ses analyses que l'auteur était âgé, conservateur et un Homme. Ce qui nous avance bien puisque c'est la description type d'une bonne part de l'Administration Trump !

Reste à trouver de qui il s'agit sachant que l'auteur a volontairement brouillé les pistes en utilisant des tournures de différents membres de l'Administration Trump !



Moteurs de recherche

Les moteurs de recherche se sont développés avant que l'IA devienne « mainstream » mais ils font de plus en plus appel à l'IA pour améliorer leurs fonctionnalités.

L'IA joue notamment un rôle clé dans la recherche d'images, pour les tagger automatiquement ou pour rechercher des images similaires. Cette dernière fonction s'appuie de plus en plus sur du deep learning et des réseaux de neurones convolutionnels.

Les outils de traitement du langage naturel sont aussi mis en œuvre pour comprendre le contenu et le contexte des recherches et pour décoder la voix des vidéos.

Google utilise depuis quelques années son outil maison RankBrain pour améliorer la pertinence des recherches, en complément de l'historique PageRank. Il serait utilisé dans plus de 15% des requêtes d'utilisateurs. Le système détermine les termes qui ont un sens voisin des mots utilisés dans la recherche en fonction de son contexte.

Antidot (1999, France, $3,5M) est connu pour son moteur de recherche pour entreprises. Il propose une fonction de classification automatique de contenus ainsi que d'amélioration de la pertinence des résultats de recherche s'appuyant sur du machine learning.

Sinequa (2002, France, $5,33M) est un fournisseur de solutions de big data et d'analyse de données pour les grandes entreprises. Il fournit un moteur de recherche sémantique capable d'exploiter les données issues de nombreux progiciels (ERP, CRM, gestionnaires de contenus, etc). La société a annoncé en 2015 investir dans le machine learning pour améliorer la performance de ses solutions.

Il existe aussi de nombreux moteurs de recherche spécialisés comme pour les métiers juridiques, vus plus loin dans la partie correspondante.

Dans le domaine de la recherche, nous avons notamment **Iris.ai** (2015, Norvège, $422K), qui facilite la recherche documentaire de travaux de recherche et affiche des nuages de mots clés facilitant la navigation dans les résultats.

Elle ambitionne aussi d'automatiser certaines fonctions des chercheurs sur le plus long terme (vidéo). La startup organise aussi des Scithons, des hackathons scientifiques permettant la mise en réseau de chercheurs et d'entreprises.

Générateur de textes et de résumés

La génération de texte à partir de données brutes issues de bases de données ou de résumés à partir de textes est un autre pan du marché[407]. Nombre de ces solutions sont exploitées dans les médias comme nous le verrons dans la partie correspondante.

Le franco-américain **Yseop** (2008) est un de ces spécialistes. Basé à Lyon et à Dallas, il propose notamment Savvy, un plugin pour Excel qui traduit en texte compréhensible les données d'un graphe. Les techniques employées associent un moteur de règles et des algorithmes génétiques. Il a un concurrent américain, avec l'outil Wordsmith d'**Automated Insights** (2007).

[407] Sur la génération de résumés, voir A Deep Reinforced Model for Abstractive Summarization de Romain Paulus, Caiming Xiong, et Richard Socher, novembre 2017.

Narrative Science (2010, USA, $43,4M) est ainsi capable de rédiger tout seul des textes à partir de données structurées quantitatives et non structurées, avec son outil Quill. Il est utilisé dans les médias et dans le marketing. C'est un peu un équivalent des solutions du français Yseop. L'un des usages typiques est de produire une brève d'information sur le cours de la bourse ou les résultats trimestriels d'une société. Quill permet notamment d'exploiter les données issues de Google Analytics ou d'historique de transactions financières. La startup a été créée par un ancien de Google et de Carnegie Mellon.

C'est une information dont le formatage est très répétitif. La startup vise les marchés de la distribution, financiers et les services publics. La société complète depuis 2016 les textes qu'elle génère avec des graphes générés par la startup **Qlik** (1993, USA).

Récital.ai (2017, France) s'est attelée au problème de la gestion des mails répétitifs pour faire gagner du temps aux cadres dans les entreprises. Leur solution exploite des exemples de mails échangés dans l'entreprise. Elle classe automatiquement vos mails par type de catégories et thèmes, et identifie l'intention et les questions de l'émetteur. Le logiciel cherche la réponse dans le système d'information de l'entreprise, via des Api (SalesForce, ect) ou déclencher un processus (reporting, autre). Il propose alors une réponse et va la placer dans les brouillons. Si vous êtes OK, vous pourrez envoyer la réponse, quitte à la modifier avant. Manque de bol, pour un consultant indépendant comme moi, c'est moins évident, vu que mon système d'information est des plus limités !

De nombreuses startups sont positionnées sur ce secteur, comme **Arria NLG** (2011, UK, $40,3M) qui vise les marchés financiers, des utilities, de la santé et du marketing, les français **LabSense** (2011, France) et **Syllabs** (2006, France, $2M), **Articolo** (2014, Israël) qui produit des articles automatiquement, **Retresco** (2008, Allemagne) qui produit automatiquement des comptes-rendus de compétitions sportives et **Textomatic** (2015, Allemagne) ou **Automated Insights** (2007, USA, $10,8M).

Enfin, citons côté applications grand public, le cas de **Google Smart Compose**, présenté lors de Google I/O en mai 2018, qui remplit les mails automatiquement dès que l'on tape quelques mots (vidéo).

Robotique

La robotique est un domaine à part entière qui tire de plus en plus partie des briques techniques de l'IA. Au même titre que l'IA, la robotique est une science intégrative qui associe capteurs, mécanique, batteries et logiciels.

La notion de robot remonte à l'antiquité mais le mot serait apparu en 1920 dans une pièce de l'écrivain tchèque Karel Capek. Le premier robot mobile capable de s'adapter à son environnement était **Shakey** (1966-1972). Il était équipé de divers capteurs dont une caméra et des détecteurs de proximité et relié à des mini-ordinateurs DEC PDP-10 et PDP-15 via une liaison radio.

Classes de robots

Les définitions d'un robot ont évolué avec le temps. Aujourd'hui, on évoque un engin interagissant avec le monde physique pour accomplir diverses tâches et qui s'adapte à l'environnement.

Voici une gradation de la notion d'automate et de robot de mon cru :

- **Automate** : il répète à l'identique un geste programmé, via un logiciel ou par la saisie d'un geste humain. C'est là que l'on peut ranger les machines d'usinage à commande numérique, les robots de peinture qui exécutent systématiquement le même geste ainsi que les imprimantes 3D. Les robots de chirurgie télécommandés sont aussi dans cette catégorie.
- **Robot** : qui ajoute à l'automate la capacité à réagir à son environnement avec des règles programmées de manière traditionnelle par logiciel. C'est le cas d'un robot d'embouteillage qui sait s'arrêter si un incident est détecté par des capteurs simples. Les premiers robots de cette catégorie ont été créés par Unimation et installés chez GM en 1961. De nombreux robots industriels manipulateurs ont été créés pendant les années 1970, aux USA (Cincinnati Milacron, Unimation), au Japon (Hirata) et en Suède (ASEA).
- **Robot** : qui réagit à son environnement grâce à des sens qui font appel à de l'intelligence artificielle et notamment la vision. C'est le cas de nombreuses catégories de drones et de certains robots humanoïdes. Cette catégorie de robots évolue donc en liaison étroite avec les progrès récents de l'IA notamment dans l'apprentissage profond.
- **Robot** : qui en plus des fonctions précédentes est doté de capacités d'apprentissage et d'adaptation. Ils sont plutôt rares.

Les robots sont souvent dédiés à des tâches dangereuses (centrales nucléaires, déminage), répétitives (peinture), stressantes (assemblage en usine), fatigantes (manutention, BTP, tonte de la pelouse), ennuyeuses (vissage), répugnantes (nettoyage) ou impossibles à réaliser de manière classique (rovers sur Mars, drones aériens). Ils interviennent aussi là où ils sont moins chers dans la durée que des opérateurs humains.

La robotique nécessite l'intégration de très nombreuses disciplines : la mécanique, les moteurs, les capteurs et les sens artificiels (vision, toucher, ouie, gaz, humidité, pression, température, proximité), la planification et le raisonnement.

Un robot est composé de très nombreux agents qui doivent être bien coordonnés. Il doit accomplir des tâches avec plus ou moins de degrés de liberté et d'initiative. Il doit pouvoir s'adapter à son environnement et gérer les imprévus. Et enfin, il doit respecter les fameuses lois de la robotique de l'écrivain Isaac Asimov issues de "I, Robot" (1950)[408].

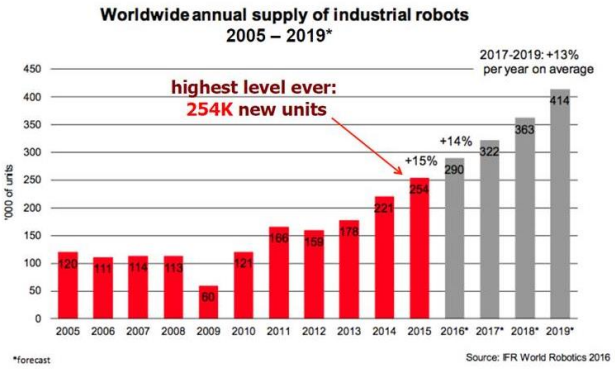
Les sciences de la robotique cherchent à répondre à de nombreuses questions clés :

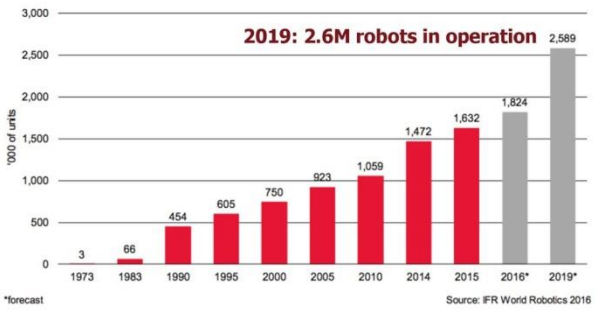
- Comment le robot peut-il se représenter le monde qui l'entoure ? C'est une question d'interprétation de ses sens visuels et autres.
- Comment doit-il réagir aux événements qu'il perçoit ?
- Comment peut-il apprendre de l'expérience ? Comme lorsqu'il apprend à éviter un obstacle de manière préventive et non pas au dernier moment.
- Comment doit-il interagir avec l'utilisateur ?
- Comment équilibrer ses objectifs et les contraintes de son environnement ?
- Comment peut-il planifier ses tâches ?

[408] Un robot ne peut porter atteinte à un être humain, ni, en restant passif, permettre qu'un être humain soit exposé au danger. Un robot doit obéir aux ordres qui lui sont donnés par un être humain, sauf si de tels ordres entrent en conflit avec la première loi, et un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.

Marché des robots

Selon **IDC**[409], le marché de la robotique mondial était de $71B en 2015 et devrait atteindre $103B en 2018 et $135B en 2019, générant une croissance annuelle de 17%. Le marché des services en robotique était estimé de son côté aux alentours de $9B à $11B en 2015 selon les sources. Il s'est vendu presque 300 000 robots industriels en 2016 et c'est la Chine qui capte la plus grande partie de ce marché, tant en production qu'en installations.

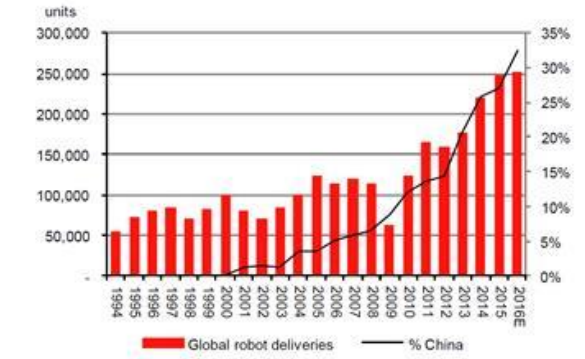




Les gros consommateurs de robots sont sans surprise les grands pays industriels : la Chine, la Corée du Sud, le Japon, les USA, l'Allemagne et l'Italie, qui est devant la France. La base installée des robots industriels serait d'environ 2 millions d'unité en 2017.

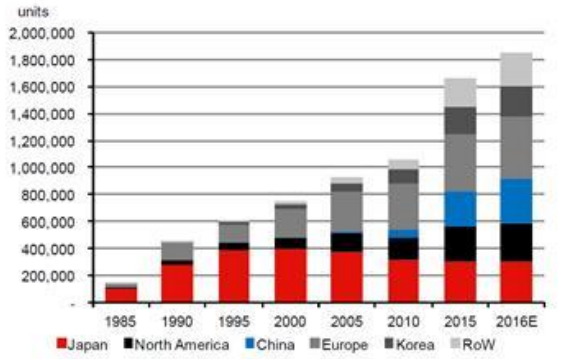
Le marché est surtout concentré sur les robots industriels mais ce sont les robots humanoïdes qui font le plus parler d'eux. Les japonais cherchent depuis des décennies à conçevoir de tels robots capables de s'occuper de leur population vieillissante. C'est un choix technologique lié à un choix politique de ne pas favoriser l'immigration. La population japonaise est d'ailleurs en déclin du fait d'un faible taux de natalité.

Fig 16 China robot demand as a % of total



Source: IFR, Macquarie Research, September 2016

Fig 17 China has driven 36% of the installed base growth between 2010 and 2016E amongst key

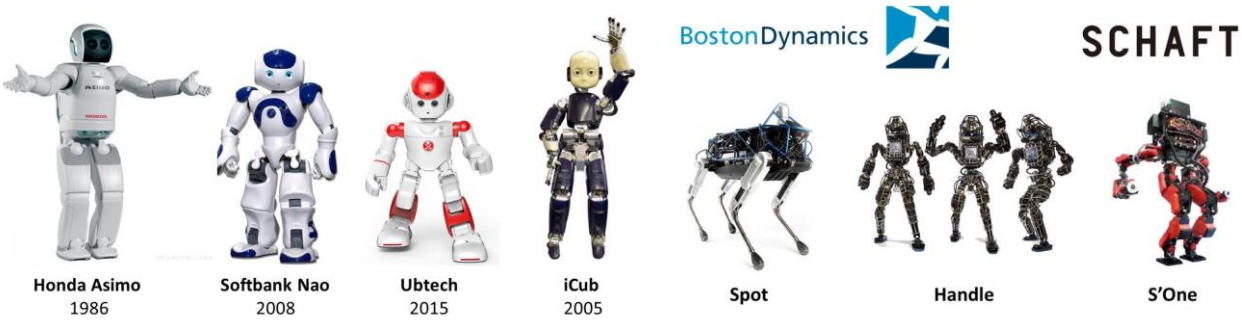


Source: IFR, Macquarie Research, September 2016

Robots humanoïdes

Le robot humanoïde le plus avancé du côté de sa capacité à se mouvoir est probablement Asimo de **Honda**, créé en 1986 et régulièrement mis à jour depuis. Il danse, court, monte et descend les escaliers et peut aussi tourner en rond. Sa dextérité est par contre moyenne et il n'est pas très fiable. C'est toujours un engin de laboratoire et de démonstration, dont les versions succesives sont généralement construites à une douzaine d'exemplaires. Malheureusement, Honda a abandonné le projet Azimo en juin 2018 ! Au même moment, le Japon décidait finalement de favoriser une immigration de travail. Les robots ne sont pas encore prêts à remplacer les Humains !

[409] Voir The Multi-Billion Dollar Robotics Market Is About to Boom, 2016.



Cette soif de robots explique les investissements de Softbank, d'une part avec l'acquisition en 2013 du français **Aldebaran Robotics**, devenu **Softbank Robotics** et d'autre part, avec celle de **Boston Dynamics** et de **Schaft** auprès de Google en 2017. Les robots humanoïdes Nao et Pepper de Softbank Robotics illustrent l'état de l'art actuel. Ils ont une belle capacité de mouvement grâce à une mécanique de bon niveau, surtout Nao. Ils interagissent en parlant avec l'utilisateur, mais de manière encore limitée. Pepper est censé capter les émotions des humains qu'il a en face de lui, grâce à IBM Watson, mais sa capacité de dialogue est encore approximative. Ces robots sont des SDK sur lequels sont construits des applications métiers comme un agent de renseignement pour un centre commercial, un point de vente (Softbank ou Nespresso à Tokyo) ou un lieu de transport.

Les agents conversationnels sont des chatbots vocaux qui ne répondent qu'à des questions scriptées et en fonction des bases de données ou de connaissances auxquelles ils sont connectés.

Les robots les plus impressionnants du moment n'ont aucune capacité de dialogue. Ce sont ceux de **Boston Dynamics**, Spot Mini, Handle (*ci-dessus, à droite*) et Atlas, handle étant capable de rouler avec habileté et de déplacer des paquets dans des étagères et Atlas, de faire un saut périlleux arrière et de courir en extérieur[410]. Leur capacité à comprendre leur environnement en temps réel constitue un réel progrès. Mais ce sont des prototypes, pas des produits déployés dans les entreprises.

Spot Mini est un robot-chien capable de manipuler des objets simples avec son bras télescopique. Les vidéos de ses démonstrations sont impressionnantes, notamment lorsqu'il ouvre une porte et qu'une personne l'en empêche et qu'il résiste[411]. Seulement voilà, Spot Mini est en fait télécommandé ! Son autonomie concerne la capacité à se mouvoir, pas à prendre des décisions sur la direction à prendre et le geste à accomplir ! Cela n'apparaît jamais dans les vidéos de démonstrations qui sont donc trompeuses. On peut par contre le constater dans les démonstrations publiques comme celle qui était réalisée lors de la conférence **USI 2018** à Paris (vidéo).

L'un des robots qui déffraye le plus l'actualité depuis 2017 est sans conteste Sophia de **Hanson Robotics** (2013, Hong Kong, $21,7M).

---

[410] La vidéo du saut-périlleux arrière du robot Atlas : est impressionnante mais il semblerait que ce soit un mouvement appris. Voir aussi sa course en extérieur ainsi que son agilité de coureur avec son état en octobre 2019.

[411] Voir la vidéo Testing Robustness, février 2018.

Ce robot humanoïde comprend un visage féminin animé, dans la lignée de nombreux robots japonais qui l'ont précédé. Au départ, le robot n'était qu'un buste avec un visage animé[412], puis il a gagné des jambes très rudimentaires.

Le robot était présenté par son créateur, David Hanson, à la télévision comme dans le Tonight Show de Jimmy Fallon en avril 2017 (vidéo) *ci-contre*, dans diverses conférences comme à la conférence Future Investment Initiative à Riyad en octobre 2017 (vidéo) à la suite de laquelle l'Arabie Saoudite lui octroyait la citoyenneté en novembre 2017[413], au WebSummit de novembre 2017 à Lisbonne (vidéo) ou encore au CES 2018 (vidéo).

En octobre, Sophia rencontrait le Président de l'Azerbaïdjan, Ilham Aliyev après avoir obtenu un visa à l'aéroport de Bakou à son arrivée[414].

A chaque fois, le robot est démontré au niveau de ses capacités d'agent conversationnel, capable de dialoguer avec son interlocuteur. Mais toutes ces démonstrations relèvent de l'esbrouffe. Les conversations de Sophia sont en grande partie pré-scriptées, et ne sont donc pas du tout spontanées[415] ! L'agent conversationnel de la startup doit s'appuyer sur le blockchain **SingularityNet** qui est mis au point par Ben Goertzel, le directeur scientifique de Hanson Robotics, par ailleurs un des grands spécialistes de l'intelligence artificielle générale et ardent promoteur de la singularité.

Sophia impressionne par le réalisme de son visage. Mais ce n'est pas le premier robot à jouer dans cette cour. Le plus connu du genre est le **Geminoid** du Japonais Hiroshi Ishiguro (*ci-dessous à gauche)*[416].

L'affaire a généré des émules. Un chatbot développé avec l'aide de **Microsoft Research** obtenait un permis de résidence symbolique dans le quartier branché de Shibuya à Tokyo en novembre 2017[417] ! Probablement parce qu'il n'était pas devenu raciste ?

En 2014, le robot **Bina48** était encore plus impressionnant que Sophia. Il était même utilisé pour donner des cours à l'Université Notre Dame de Namur en Californie (vidéo) ! Il a en fait été créé par David Hanson, le fondateur de Hanson Robotics à l'origine de Sophia. Il y a aussi le double de **Nadia Magnenat-Thalman** (*ci-dessous à droite*), sans compter les nombreux animatronics qui servent aux productions cinématographiques.

---

[412] A noter que la startup **Realbotix** (USA) est spécialisée dans la création de visages animés qui semblent très voisins de celui de Sophie de Hanson Robotics. Ils les installent dans des poupées féminines servant de sex-toys.

[413] Un robot citoyen ? Alors qu'il n'était même pas mobile au moment de l'octroi de cette nationnalité. C'est du pipeau politique dans un pays qui n'accorde que peu de droits civiques aux ouvriers étrangers qui travaillent chez eux et qui sont souvent maltraités. Ils sont plus de 9 millions dont seulement 100 000 sont occidentaux ! "*Saudi Arabia requires foreign workers to have their sponsor's permission to enter and leave the country, and denies exit to those with work disputes pending in court. Sponsors generally confiscate passports while workers are in the country; sometimes employers also hold passports of workers' family members*" (**Wikipedia**). Le pays a autorisé les femmes à conduire, mais seulement après avoir accordé la nationalité au robot Sophia !

[414] Voir AI Humanoid 'Sophia' Is Granted First Ever Robot Visa, Speaks With President, de Paul Amstrong, octobre 2018. La vidéo du dialogue avec le Président est surtout un long monologue de ce dernier. Le robot écoute poliment et répond. Mais la voix off Azéri couvre la voix en anglais de Sophia, il est donc impossible de saisir quoi que ce soit du dialogue.

[415] Que sait réellement faire Sophia, le robot dont l'intelligence est contestée ?, de Morgane Tual, février 2018, qui reprend les propos agacés de Yann LeCun sur le brouhaha marketing autour de Sophia. Et la vidéo Hanson Robotics Sophia is a Fake!, janvier 2018, bien complétée par Mama Mia It's Sophia: A Show Robot Or Dangerous Platform To Mislead?, de Noel Sharkey, novembre 2018.

[416] Voir Super realistic robots at Tokyo Game Show leave social media divided with a hot debate, juillet 2018.

[417] Voir AI 'boy' granted residency in central Tokyo, novembre 2017.

De nombreux robots avec des capacités mécaniques plus limitées sont proposés pour servir de centres de renseignement ambulants dans des lieux publics comme les centres commerciaux ou les aéroports.

Ce sont en quelque sorte des tablettes à roulettes, comme chez le chinois **Qihan**, le français **Hease Robotics** (2016, France), un autre robot français, créé à Lyon et animé par les logiciels d'un autre Lyonnais, la startup **Hoomano** (2014, France). Il y a aussi ceux d'autres startups françaises, **Partnering Robotics** (2007, France) et **Wyca Robotics** (2015, France) ou **Robosoft Kompaï** (1996, Japon, $15,7M).

Leur fonction principale est donc liée à l'agent conversationnel métier qu'ils intègrent et qui gère un spectre souvent étroit de discussion. Je me moque un peu de ces tablettes à roulettes mais leur forme a une utilité.

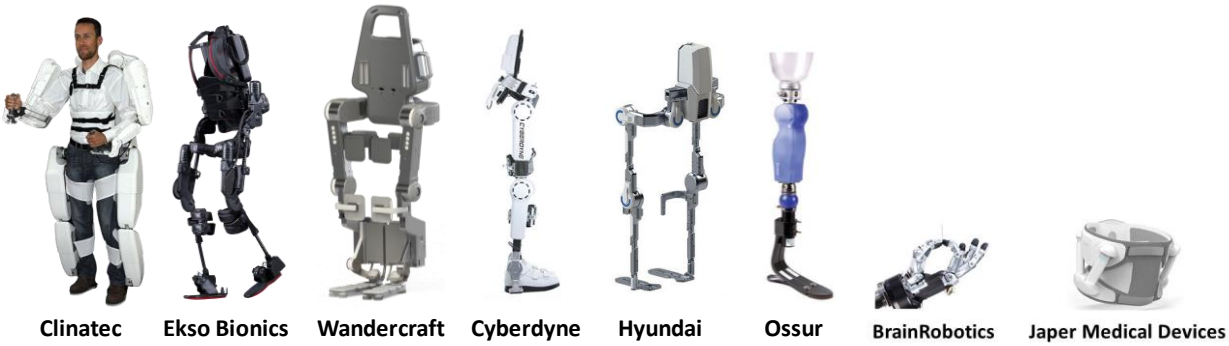
Elle est d'abord plus facile à fabriquer et gérer et elle permet d'éviter de faire basculer le robot dans la vallée de l'étrange (uncanny valley). Elle corrrespond au sentiment ressenti qui peut être désagréable lorsque l'on a à faire à un robot trop proche d'un être vivant.

Exosquelettes

De leur côté, les exosquelettes ont moins de capteurs sensoriels. Ils sont surtout pilotés par l'utilisateur, notamment via la partie de leur corps qui fonctionne encore comme pour **Wandercraft** (2013, France, $23,5M), qui démontre depuis fin 2017 son premier prototype opérationnel, ou **Cyberdyne** (2004, Japon), et qui remonte parfois au cortex moteur dans les lobes biparietaux comme pour le projet d'exosquelette à quatre membres du laboratoire grenoblois **Clinatec** qui est destiné aux tétraplégiques. **Ekso Bionics** (2005, USA, $70,8M) vise de son côté aussi bien les compensations de handicaps que les applications industrielles. Ses exosquelettes sont en tests dans des usines de Ford depuis fin 2017[418]. **RB3D** (2001, France) met au point pour sa part des exosquelettes ciblant notamment le marché de la construction.

[418] Voir Ford Testing Exoskeleton Suit For Its Factory Workers, novembre 2017.

Il y a moins d'IA et plus de mécanique dans ces produits. L'un des points clés est la miniaturisation des moteurs et des batteries pour rendre ces engins aussi légers et pratiques que possible. Les exosquelettes peuvent être partiels comme avec la main robotisée de **BrainRobotics** (2015, USA) et l'exosquelette lombaire Atlas de **Japet Medical Devices** (2016, France, $1,6M).

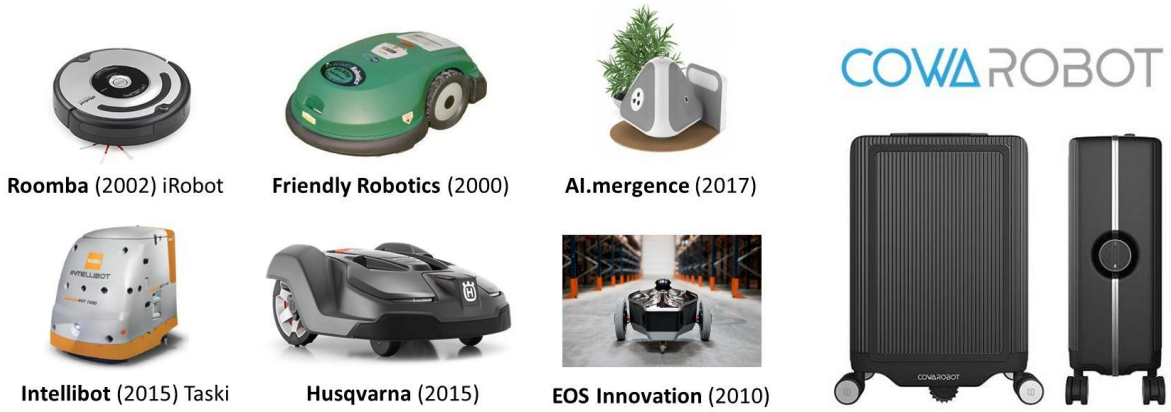


Robots de services

De nombreux robots que nous venons de voir sont inspirés d'œuvres de science-fiction. Les robots commercialisés en volume et véritablement opérationnels prennent bien d'autres formes.

C'est le cas des robots « de sol ». Qu'il s'agisse d'aspirateurs (Roomba de **iRobot**), de tondeuses à gazon (**Friendly Robotics** et **Husqvarna**) ou de systèmes de sécurité mobile (**EOS Innovation**[419], **AI.mergence**), leur principale fonction est de se mouvoir au sol, d'éviter les obstacles, de balayer optimalement une surface donnée, et de réaliser une tâche mécanique spécifique.

Leurs capteur de proximité voire visuels leur permettent de cartographier leur environnement et de s'y mouvoir. Ils doivent aussi souvent pouvoir retrouver leur station de recharge de batterie. Une variante de ce genre de produit est l'étonnante valise robot de **Cowarobot** qui suit automatiquement son utilisateur (vidéo), que j'avais découverte à l'occasion du CES 2017 et qui est maintenant commercialisée.



Les loisirs sont un autre domaine pour les robots, avec le robot joueur de ping pong du fabricant de composants **Omron** (*ci-dessous*, vu au CEATEC de Tokyo en octobre 2014 et qui n'est qu'un démonstrateur pour un fabricant de composants) ou le ramasseur de balles de tennis de **Tennibot** (2016, USA). Le joueur de ping pong robotisé exploite surtout un système de vision stéréoscopique couplé à un système prédictif de la position de la balle en fonction des gestes de son compétiteur humain. Le ramasseur de balles utilise ses capteurs de vision pour détecter les balles et les ramasser.

[419] L'ESN **Econocom** a décliné le robot d'EOS Innovation pour en faire un robot d'inspection de data-center en 2017, Captain DC. EOS Innovation était une filiale du français Parrot qui a été ensuite acquise par **Econocom**. Voir https://vimeo.com/170005575.

Omron (2014)

Tennibot (2017)

Bionic Bar (2015)

Denso Barrista Robot (2017)

Les robots d'usines sont de leur côté mis à contribution pour devenir barmen ou gestionnaires de machine à café comme avec le Bionic Bar, installé dans les paquebots du croisiériste américain **Royal Carribean** et le **Denso Barrista Robot** vus au CES 2017, une déclinaison d'un robot d'usine pour un usage grand public de démonstration, Denso étant surtout un équipementier pour l'industrie automobile.

Il existe aussi plein de robots transporteurs de charges pour les entrepots, comme les robots manutentionnaires de **Kiva Systems** (2003, USA, $17M) qui ont été acquis par Amazon pour $775M en 2012 (*ci-dessous à gauche*).

FuelMatics (2008)

Dans les transports, les robots de **FuelMatics** (2016, Suède) remplissent automatiquement votre réservoir d'essence si vous avez installé leur bouchon spécifique dans votre véhicule (*ci-dessus à droite*).

Dans l'aérospatial, les robots doivent être très autonomes. C'est le cas des rovers **Sojourner** et **Curiosity** qui explorent Mars. Les communications aller et retour entre Mars et la Terre durent plus de 45 mn. Ces robots doivent donc se débrouiller tous seuls en fonction de leur plan de charge. La conséquence est qu'ils sont plutôt lents.

Les télescopes spatiaux sont aussi très autonomes, commele **James Webb Telescope** qui sera lancé en retard en 2019 ou 2020 et mettra plus de deux semaines à se déployer avec des dizaines d'opérations de dépliement de sa structure en origami.

Sojourner (1997)

Curiosity (2012)

Space Shuttle (1981)
ISS (2001)

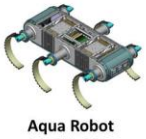
Predator, 1995)

Global Hawk, 1998

Safran Patroller, 2009

Robonaut (2011)

JWST (2019)



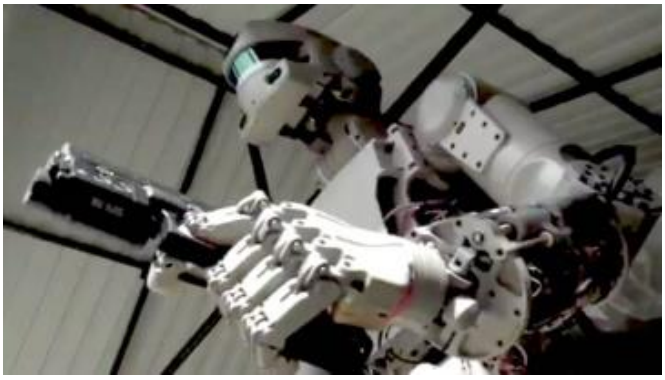
Aqua Robot
McGill, 2007

Odyssey IV
MIT, 2008

A6K
IFREMER, 2017

Les drones militaires savent gérer leur vol de manière autonome mais sont pilotés à distance, notamment pour les ordres d'observations ou de lancements de missiles. Il en va de même pour les drones sous-marins qui sont téléguidés.

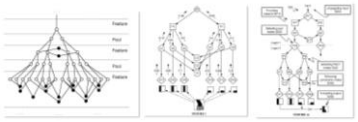
On se demande cependant à quoi peut rimer ce robot russe, le **FEDOR**, tout droit sorti de Robocop qui tire avec précision sur une cible fixe (vidéo). Si l'objectif est de faire peur sur les risques de l'IA et de la robotique, il sera rapidement atteint.



Comment fonctionnent tous ces robots ? Un peu comme pour le Machine Learning et le Deep Learning, leurs créateurs ont rarement développé des logiciels idoines pour leur donner vie.

Ils s'appuient le plus souvent sur des SDK du marché. Plusieurs startups sont présentes sur ce marché et notamment **Vicarious** (2010, USA, $122M), **Kindred** (2013, USA, $44M) et **Osaro** (2015, USA, $13,3M). Des startups comme l'Américaine **Neurala** (2006, USA, $15,2M) sont spécialisées dans l'IA pour le pilotage automatique des drones avec leur SDK Brain4Bots intégrant deep learning, vision artificielle et support de GPU comme ceux de Nvidia.

vicarious



KINDRED

OSARO

neurala

KINEMA SYSTEMS

COVARIANT.AI

On peut aussi compter sur **Kinema Systems** (2016, USA) et son système de vision 3D pour robots de manutention, Kinema Pick associé au capteur 2D/3D (vidéo) ainsi que **RightHand Robotics** (2014, USA, $11,3M) qui est sur un créneau voisin avec des solutions robotiques intégrées pour la préparation de colis dans le commerce en ligne.

Covariant (2018, USA, $7M, anciennement Embodied Intelligence) est issue de Berkeley[420] et développe des fonctionnalités permettant à des robots d'être plus versatiles dans leur gestuelle et maîtrise de la manipulation d'objets.

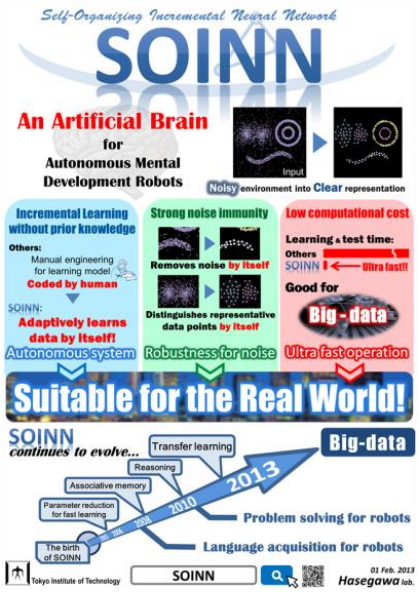
20 ans après la renaissance des réseaux neuronaux, en 2006, le japonais **Osamu Hasegawa** créait les réseaux neuronaux auto-organisés incrémentalement ("Self-Organising Incremental Neural Network" ou SOINN), utilisables dans des réseaux neuronaux auto-réplicables et capables d'auto-apprentissage.

En 2011, son équipe développait un robot utilisant ces SOINN capable d'auto-apprentissage (**vidéo**), illustrant magistralement les applications des réseaux neuronaux (*ci-dessus à droite*).

La robotique est encore un grand champ d'expérimentation et de makers. Un grand nombre de projets d'étudiants et de chercheurs tâtonnent pour faire avancer le domaine. On a par exemple des robots de tri de pièces de Lego (source) ou de résolution du Rubik's Cube en une demi seconde (vidéo).

[420] AI Startup Embodied Intelligence Wants Robots to Learn From Humans in Virtual Reality, novembre 2017. Voir la vidéo où leur robot peut faire des nœuds avec des ficelles, après avoir réussi le pliage de serviette (vidéo). Ils développent des techniques d'aprentissage des gestes par la réalité virtuelle (vidéo).

Divers autres robots originaux qui n'ont pas forcément vocation à être industrialisés et commercialisés. Vous avez ce projet de robot menuisier domestique développé par le **MIT** (vidéo, février 2018), ce robot abeille de pollinisation développé au Japon par le National Institute of Advanced Industrial Science and Technology[421] ou le robot **Salto-1P** de Berkeley qui peut bondir pour traverser une pièce (vidéo)[422].



De nombreuses startups se lancent avec plus ou moins de bonheur dans le domaine. Les me-too sont légion et on attend toujours des robots capables de bien appréhender leur environnement et d'interagir avec. Se développe également une économie de services et d'ingénierie en robotique[423].

Cela explique pourquoi nombre de métiers manuels sont bien plus protégés aujourd'hui que le sont certains métier de cols blancs gérant des processus répétitifs, bien plus faciles à automatiser.

Le marché de la robotique est un secteur d'activité très difficile, à l'instar de celui des objets connectés. C'est un défi rarement relevé que d'être bon à la fois dans le matériel et le logiciel. Les sociétés développant la partie matérielle des robots sont souvent déficientes en logiciel où elles n'investissent pas assez et où l'état de l'art n'évolue pas assez rapidement. Ou alors, les robots ne sont que des tablettes roulantes qui vont s'appuyer ensuite sur les grandes plateformes de chatbots vocaux pour dialoguer avec des utilisateurs.

D'où le fait que nombre de robots commercialisés sont très spécialisés comme dans l'agriculture ou dans les usines ou pour la manutention dans les entrepots. Le robot généraliste est un projet très long terme ! C'est au minimum un gros SDK qui doit donner lieu au développement de nombreuses IA étroites pour son ou ses usages.

Ces écosystème lent à créer et comme dans les smartphones, il n'y a pas de place pour un très grand nombre d'acteurs. Soit le marché est fragmenté est les écosystèmes ne se développent pas, soit il se consolide au prix de l'élimination d'une bonne part des petits acteurs. Dans ce cadre, il était intéressant de visiter le stand « d'écosystème » de Pepper chez Softbank Robotics à VivaTech.

Les Français qui se lancent dans la création de startups en robotique tombent donc dans l'un des deux cas de figure. Soit des robots spécialisés, soit des robots généralistes, difficiles à mettre au point.

---

[421] Voir Japan has just invented Robo-bees that can legitimately pollinate the earth, février 2017.

[422] Voir Insolite : Le robot le plus petit au monde qui fait des sauts de géant, de Loïc Bremme, juillet 2918.

[423] Comme avec **iSee Automation**, une startup issue du MIT qui est financée dans le cadre du fonds de deep techs **The Engine** lancé par ce dernier et doté de $200M (source).

TwinswHeel (2016, France) créé des robots autonomes de livraisons dédiés aux environnements fermés comme des usines, semi-fermés et ouverts (vidéo). Ils ont déjà exposé au CES et été soutenus par la Poste. C'est le modèle de livraison du « last mile », le point clé étant ensuite celui des derniers mètres pour arriver entre les mains du destinataire ! Le robot qui a une charge utile de 40 kg sait bien entendu éviter les obstacles le long de son trajet. Il sait aussi traverser la rue aux passages cloutés et interpréter la signalisation. Il roule entre 6 km sur le trottoir et 30 km/h sur la route[424].

Cybedroid (2011, France) et son robot Alice qui rappelle Pepper, mais sans l'écran. Il a des mains et se déplace sur roulettes. Ses moteurs lui procurent 20 degrés de liberté et il dispose de huit heures d'autonomies pour déambuler dans des lieux publics et servir de concierge d'entreprises et accueil. La valeur est dans l'agent conversationnel métier qui motorise le robot !

NeedaBot (2016, France), anciennement HoCaRo, propose un robot de soutien aux aides soignants dans les EHPAD. Bras, tête et roulettes. Détection de chute.

Events Bots (2016, France) a développé Hope, un robot d'accompagnement d'enfants malades en milieu hospitalier, qu'il occupe pendant les soins.

Spoon (2015, France) propose un robot manipulateur d'objets (vidéo), qui est plutôt un automate, vu plusieurs fois à VivaTech et au CES. Il permet des interactions avec le public, avec une souplesse lui donnant un côté « animal ».

Le reste est dans l'écran, comme c'est le cas dans Pepper, et on retombe dans le défi logiciel et des agents conversationnels. Ses fondateurs sont issus d'Aldebaran Robotics et avaient contribué à la création de Nao et Pepper.

Cobots

Les robots collaboratifs sont utilisés à proximité d'opérateurs pour accompagner leurs gestes grâce à de meilleures facultés d'adaptation et une bonne gestion des contraintes de sécurité. Les principaux fabricants de cobots sont à la fois des entreprises récentes comme **Universal Robots** (2005, Danemark, acquis par Teradyne en 2015), **Rethink Robotics** (déjà cité), **Micropsi** (2014, Allemagne, $9,5M, avec son robot Mirai), des fabricants industriels comme **Kuka** (qui appartient maintenant au Chinois **Midea**), **ABB**, **Fanuc** tout comme quelques startups françaises telles qu'**IsyBot** (2016) ou **MIP Robotics** (2015, France, $2,3M).

Marketing et vente

Le marketing et la vente, surtout en ligne, sont l'un des marchés les plus florissants des applications de l'IA. Elles sont mises en œuvre dans toutes les étapes du cycle de vente et l'offre de startups y est tellement abondante que cela en devient risible, la cartographie *ci-dessous* totalisant 5000 logos.

Toutes ces startups ne font pas appel à de l'IA même elles sont nombreuses à s'en vanter ! L'excellent Panorama des solutions d'intelligence artificielle pour le marketing publié par Fred Cavazza en octobre 2017 permet d'y voir un peu plus clair (*ci-dessous* à droite).

---

[424] La startup n'est évidemment pas seule sur ce marché qui intéresse aussi les grands du ecommerce. Voir par exemple Alibaba mise sur un robot autonome pour les livraisons, de Pierrick Labbé. Leur G Plus roule à 15 km/h mais est positionné pour un usage dans les entrepots. Dans le même ordre d'idée, voir OpenAI Demonstrates Complex Manipulation Transfer from Simulation to Real World, de Evan Ackerman, juillet 2018, avec de l'apprentissage de geste par tâtonnement et injection d'aléatoire.

Dans le marketing amont et le planning, l'IA aide à segmenter ses clients, à comprendre leur besoin, à définir des marchés et clients cibles et à interagir directement avec eux. Le profiling d'utilisateurs dans les réseaux sociaux permet de faire du micro-ciblage d'offres.

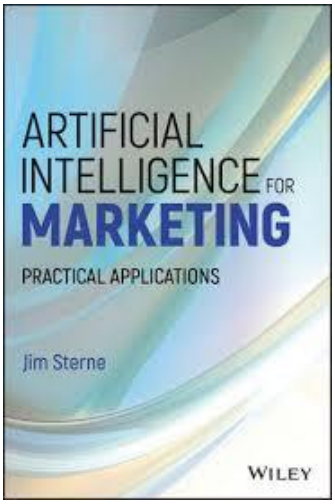
L'étape de développement de la notoriété tire parti de solutions qui aident à optimiser le plan média et le reach de ses campagnes. Les chatbots interviennent aussi bien en avant-vente qu'en après-vente et leur offre est abondante comme nous l'avons vu précédemment.

Bref, la panoplie des outils d'IA en marketing et vente est vaste, surtout pour les sites de vente en ligne.

Quel est l'impact de ces outils sur les métiers du marketing ?

Dans la majorité des cas, ils constituent une boîte à outils étendue pour les marketeurs comme l'ont été de nombreux outils de productivité depuis l'invention de la bureautique et du tableur.

L'impact peut être plus important dans le cas des chatbots car ils peuvent entraîner une substitution partielle des tâches vis-à-vis d'agents de centres d'appels, surtout pour les appels entrants et le support technique en appels entrants.

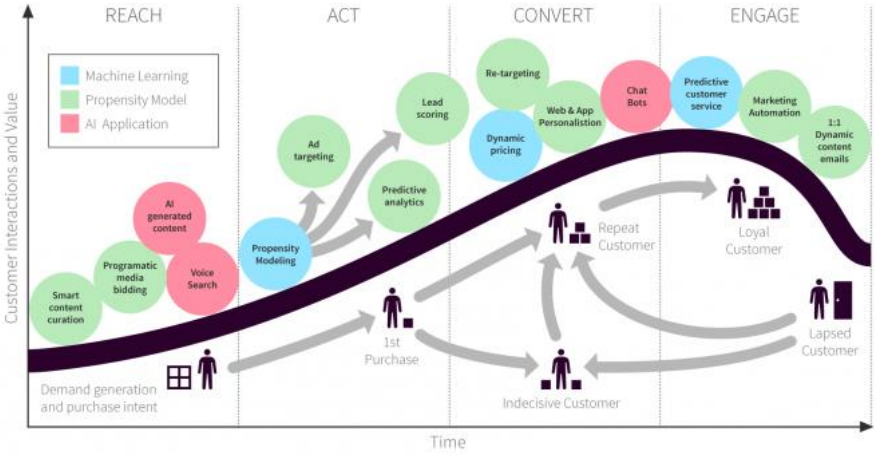


Artificial Intelligence for marketing, 2017 (361 pages[425]) est un ouvrage librement téléchargeable qui contient un excellent panorama des applications de l'IA dans le marketing.

[425] Téléchargeable ici : http://www.shabakeh-mag.com/sites/default/files/files/attachment/1396/05/1502457644.pdf.

planning	notoriété	considération	évaluation	achat	recommandation
segmentation mix marketing pricing	chatbot plan média analyse d'image SEO	chatbot optimisation site web	chatbot	chatbot recommandation upsell / cross-sell	analyse sentiments identification influenceurs

l'IA intervient dans tout le cycle de vente !



Planification

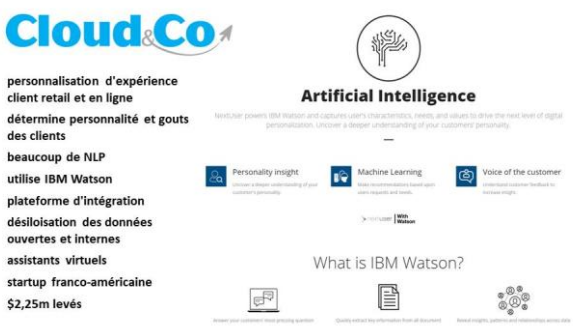
Une bonne part des startups de ce secteur proposent des solutions généralistes intégrant plusieurs outils et accédant à des sources de données externes. Elles servent à segmenter son audience, définir les bons messages et les bons canaux de communication pour les toucher.

Lucy d'**Equals3** (2015, USA, $7M) s'appuie sur IBM Watson pour segmenter les clients, définir ses messages et optimiser son média planning (vidéo).

Albert de la startup **AdGorithms** (2010, IPO en 2015) intègre des outils de segmentation d'audience, d'achats médias, d'optimisation plan média cross-channel, de tests et optimisation et d'analytics. C'est une grosse boîte noire exploitant de nombreuses sources de données. Albert s'appuie aussi sur des briques logicielles d'IBM Watson.

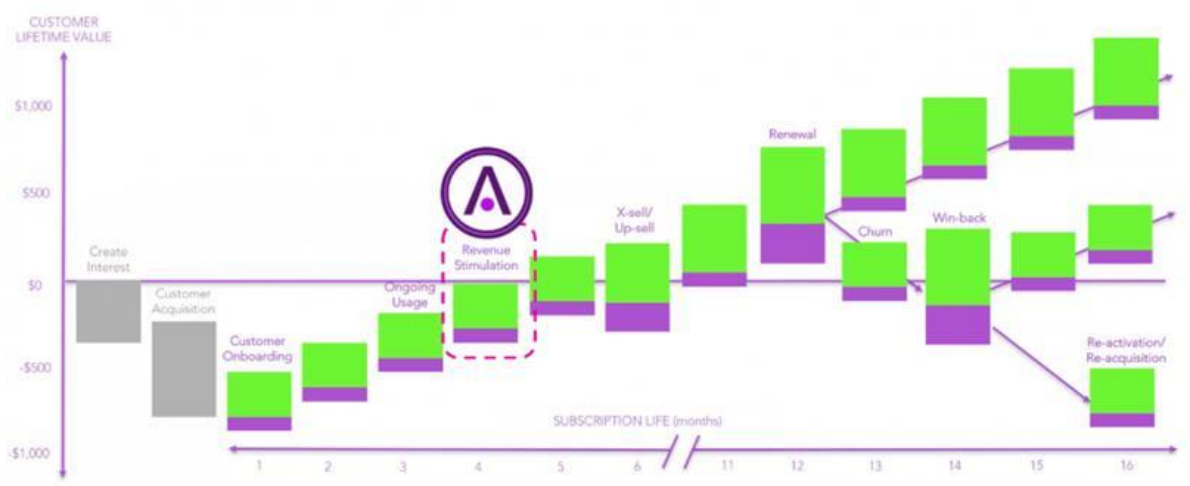
Optimove (2009, USA, $20M) aide aussi les équipes marketing à bien connaître leurs clients et à les segmenter et, dans sa boîte à outils, propose un moyen d'optimiser son mix marketing.

L'optimisation des messages et contenus est aussi le domaine des startups américaines assez bien financées que sont **Captora** (2012, USA, $27,3M) et **Persado** (2012, USA, $66M).



La planification des messages et des médias s'appuie sur la gestion et l'analyse des données issues des médias sociaux comme avec **Cortex** (2014, USA, $500K) qui prédit la réaction des Internautes aux contenus et **SimpleReach** (2010, USA, $24,2M).

Amplero (2016, USA, $25,6M) aide à ajuster son mix marketing pour générer la plus grande "customer lifetime value" (valeur générée par le client dans la durée). C'est l'outil rêvé pour optimiser ses plans marketing cross-médias. La startup présente quelques études de cas d'optimisations breadth vs depth dans la portée d'actions marketing, cross-sell vs upsell, bref pour gérer les choix cornéliens classiques du marketing. **Dynamic Yield** (2011, Israël/USA, $77,3M) a l'air d'être positionné d'une manière similaire pour optimiser un mix média.



Création

D'autres startups sont spécialisées dans la création de contenus avec une assistance de l'IA, souvent très peu décrite dans sa nature exacte.

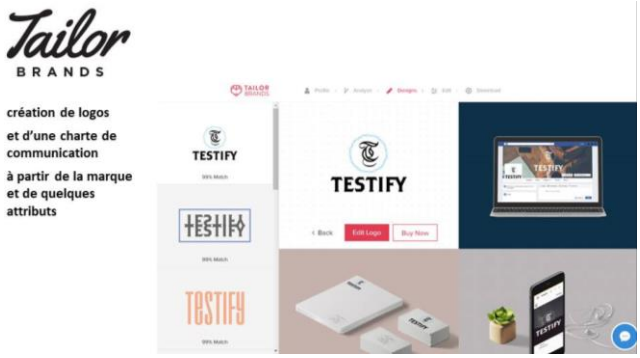
Le deep learning est utilisé dans la création graphique ou la création de sites web. C'est ce que réalise le site de création de sites web **Wix** (2006, Israël, $58,5M, vidéo). Il s'est entraîné en analysant le style de nombreux sites web et leur performance (SEO, trafic) pour en déduire empiriquement les paramètres d'un bon site, le tout, en fonction de plusieurs milliers de professions (**source**). L'outil de création de sites web va également exploiter l'ensemble des textes et visuels disponibles sur la société pour bâtir une maquette de site. La création du site est guidée par l'utilisateur qui va tout de même choisir des attributs de sa marque pour influencer le design généré et ses couleurs dominantes. Le danger de ces outils ? Ce n'est pas une surprise : il risque de faire converger les sites créés vers un design voisin.

Une approche voisine est proposée pour l'instant de manière expérimentale par **Adobe**[426]. Le tout s'intègre dans l'Adome Experience Manager CMS (outil de création de site) et exploite Adobe Sensei, la plateforme logicielle qui intègre les briques d'AI de l'éditeur. Cette IA est focalisée sur l'ajustement des contenus graphiques, pour mieux cadrer les illustrations.

[426] Voir Adobe Is Building An AI To Automate Web Design. Should You Worry?, mars 2017.

La création à base d'IA porte aussi sur celle de logos. **LogoJoy** (2016, USA, $900K) permet d'en créer avec une méthode un peu bourrine : l'utilisateur sélectionne d'abord cinq styles qui lui plaisent, indique le nom de sa société puis son slogan et enfin, choisi une couleur et cinq icones qui correspondent à son service. A partir de là, le système génère un logo. Ce n'est pas très créatif puisque le logo associe le nom de la société, son slogan et tout ou partie des icones sélectionnées.

Tailor Brands (2014, Israël/USA, $20,6M) aide aussi à créer son logo et sa charte de communication après avoir fourni quelques attributs de sa marque comme son nom et son secteur d'activité[427]. Ce n'est pas bien sophistiqué mais ça marche. La dose d'IA dedans ? Pas évidente. La startup vise le marché de volume des PMEs qui ne peuvent pas de payer une agence de création de logo traditionnelle. Le tarif est au mois !



Il y a aussi **Designhill** (2014, Inde) qui fait la même chose (vidéo) mais peut aussi s'appuyer sur des créatifs humains, là où ils ne sont pas chers, pour des tâches de création plus sophistiquées que la création de logos. Ici, le prix de la création de logos via l'IA est fixe, compris entre $15 et $150. Et on obtient un résultat en une minute.

L'agence de communication **McCann** de Tokyo annonçait en 2016 utiliser une IA baptisée AI-CD β comme directeur de création[428]. Cette IA était construite dans le cadre du « creative genome project" de l'agence. Elle aurait créé un script de publicité TV pour vendre des dragées à la menthe de la marque Clorets du groupe britannique Cadbury. Au bout du compte, celui-ci était de moins bonne qualité que celui du créatif humain de l'agence. AI CD β est matérialisé par un petit système qui dessines des idéogrammes. Il en est allé de même en septembre 2016 pour des publicités réalisées par **Burger King** avec une IA générant automatiquement le scénario (vidéo)[429].

Visiblement, l'IA en question analysait une base de 10 ans publicités TV bien labellisée avec les études d'impact pour identifier leurs éléments d'efficacité.

Cette base était croisée avec les éléments du brief de la publicité comme l'audience visée, l'objectif de la campagne et le « claim » de la campagne. L'IA faisait des recommandations sur les éléments de l'annonce comme le contexte, la tonalité musicale ou le type de célébrité à y intégrer. Le reste était réalisé par des collaborateurs de l'agence. Bref, en guise d'IA, nous avions une application de machine learning multicritère.


Human Beats AI CD in McCann Japan's Creative Battle

Share Tweet More

By Erik Oster on Sep. 1, 2016 - 10:27 AM Comment

After introducing its A.I. CD early this spring, McCann Japan decided to bit the AI-CD β robotic creative director against its human counterpart, namely creative director **Mitsuru Kuramoto**, in a creative battle. Both were given the task of creating a spot for Mondelez Japan brand Clorets Mint Tab, communicating the brand message of "instant, long-lasting refreshment that lasts for 10 minutes" and then turning to a nationwide poll to declare the winner.

En octobre 2017, l'agence Belge **DDB** faisait jouer le rôle de juré à une IA développée avec Microsoft pour sélectionner la meilleure publicité dans la dixième édition de leurs MIXX Awards.

---

[427] Voir une description du processus dans Avec l'IA, la génération de logos passe au low-cost de Lélia De Matharel, décembre 2017.

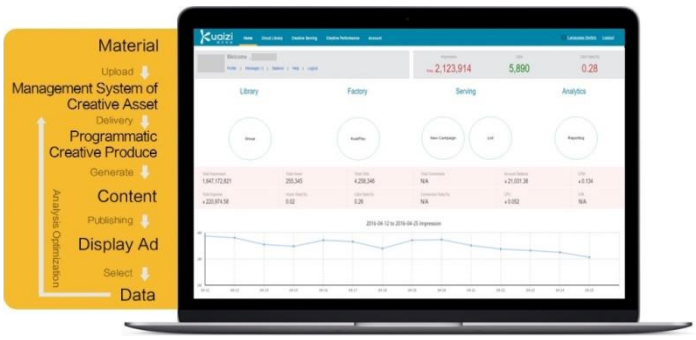
[428] Voir Human Beats AI CD in McCann Japan's Creative Battle, septembre 2016. Voir cette vidéo qui explique l'histoire.

[429] Voir Burger King's AI-written ads are beautiful disasters, septembre 2018.

Le Video Indexer Service de Microsoft servait à analyser dix ans d'archives publicitaires taggées pour prédire lesquelles seraient en haut du panier[430].

Kuaizi Technology (2013, Chine, $1,5M) propose la création de contenus basée sur l'analyse de la performance de contenus anciens. La société était présente sur le stand de LVMH de Viva Technology en mai 2018.

C'est un partenaire de Facebook. Leur « Kuaizi PCP Programmatic Creative Platform » permet notamment de faire de l'A/B testing de créations sur Facebook. Leur SEM Smart Matching System of Landing Page sert à optimiser la page d'accueil d'un site web. Probablement à base de machine learning.



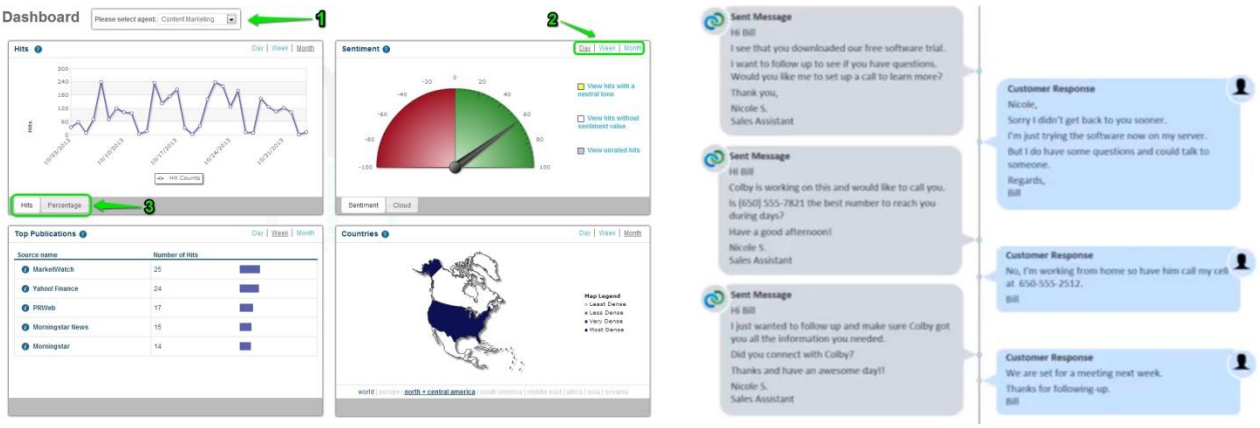
Web et analytics

Nous avons aussi des solutions d'optimisation de sites web comme **Webpage.ly** (2015, Canada) qui est focalisé sur le référencement naturel (SEO) et fonctionne en mode cloud. **Tilofy** (2013, USA, $1M) est de son côté une solution solution de prévision des tendances dans le fashion et les usages, dont les méthodes et techniques ne sont pas précisées mais qui doit faire appel à des techniques d'analyse du langage (NLP).

Network Insights (2006, USA, $77,2M, acquis par American Family Insurrance fin 2017) propose Audience.ai, un outil d'analytics qui exploite les traces des utilisateurs dans les réseaux sociaux pour définir des profils de prospects et clients ultra-précis et des messages associés ultra-ciblés. Ce n'est pas sans rappeler les méthode de Cambridge Analytica qui avait joué un rôle pour cibler les messages pro-Trump dans les *swing states* pendant la présidentielle américaine de 2016 et défrayé ensuite la chronique au printemps 2018 lorsque les méthodes associées ont été dévoilées au grand public. Cette approche marketing consistant à faire du micro-ciblage multi-factoriel sur des clients avant de lancer ses campagnes marketing est aussi proposé par une startup de Palo Alto, **Mariana** (2014, USA, $4M).

Meltwater (2001, USA, $60M en dette) propose des solutions en cloud de veille et d'analyse de l'information sur les médias en ligne et sociaux. Elle croit rapidement par acquisitions, avec quatre acquisitions en 2017 dont celle de **Wrapidity** (2015, UK, £50K), issue d'un projet de l'Université d'Oxford. Meltwater couvre la veille stratégique, la pige média en ligne, le ciblage de journalistes, l'e-réputation, l'analyse des réseaux sociaux et de sentiments sur les marques et la mesure de performance des campagnes marketing en ligne. Le tout est présenté dans un tableau de bord (*exemple ci-dessous à gauche*).

[430] Voir Meet Pearl – an AI judge who is already dishing out awards, octobre 2017.



Conversica (2007, USA, $72M) est un outil d'automatisation de la communication par mail à des prospects (*exemple ci-dessus à droite*). Comme d'habitude, les techniques d'IA ne sont pas précisées mais relèvent certainement de combinaison d'outils de NLP (traitement du langage).

L'outil s'interface comme il se doit avec ceux de SalesForce et met le client en relation avec un véritable conseiller commercial si nécessaire. C'est une sorte de Julie Desk appliquée aux premiers traitements d'une demande d'un prospect. Une explication du processus dans cette vidéo.

Spella (2014, France/Belgique) collecte et structure les données issues des commentaires des internautes pour identifier les signaux forts et faibles avec un bon niveau de granularité.

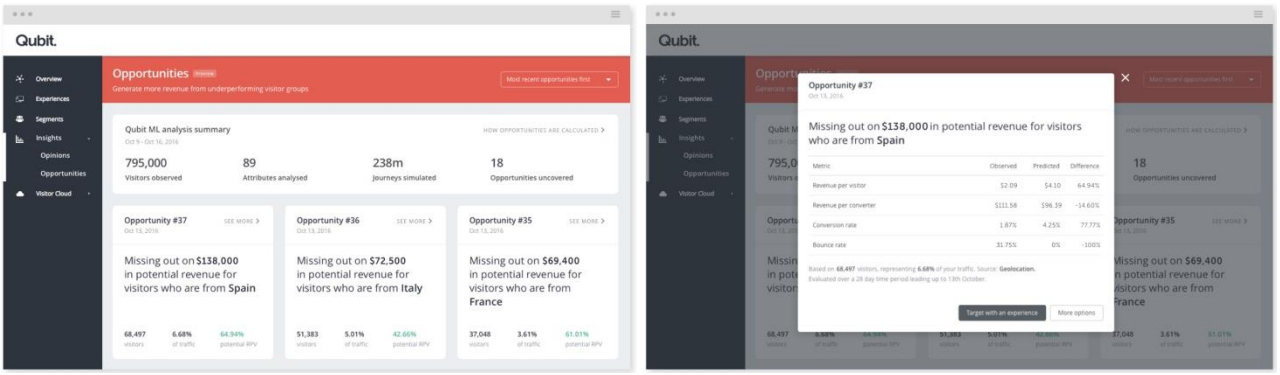
Q°Emotion (2015, France, 614K€) propose un outil d'évaluation des émotions en temps réel dans les avis clients postés sur tous les canaux sociaux. Il est notamment utilisé par la Fnac.

Support aux ventes

En complément du précédent, **People.ai** (2016, USA, $37M) fournit de son côté des outils d'aide et de coaching de conseillers commerciaux pour les former à closer les deals, par analyse de patterns d'appels antérieurs, et aussi pour les éviter de passer trop de temps avec les clients impossibles à closer. Leur outil analyse plus de 90 sources d'information dont les outils de communication de l'entreprise (mail, vidéoconférence, etc) pour alimenter le CRM. Le machine learning analyse les textes, les émotions et les activités pour identifier des points à régler dans le cycle de vente afin de l'accélerer. La startup propose aussi des vidéos de formation pour conseillers commerciaux.

Qubit[431] (2010, UK, $74M) a développé un moteur de détection automatique des opportunités de revenu à base de machine learning, exploité principalement par les sites de vente en ligne. C'est en fait un outil de segmentation automatique de clients pour identifier ceux qui sont les plus prometteurs. Le système permet aussi de piloter des campagnes en ligne d'A/B Testing d'offres commerciales ciblées.

[431] Il est dommage que cette société utilise une dénomination liée à l'informatique quantique, qu'elle n'utilise pas du tout. Quand on cherche des vidéos les concernant, on tombe bien évidemment en premier sur des explications sur les qubits de l'informatique quantique et pas sur leurs propres vidéos. C'est bien ballot !



LeadGenius (2011, USA, $19M) propose une solution intégrée de génération, de qualification et de conversion de leads en mode b2b. Elle s'appuie sur le scrapping de sources Internet diverses (LinkedIn qui contient plus de 7,5 millions d'entreprises aux USA, Yelp, AngelList et Crunchbase pour les startups, déclarations de résultats auprès de la SEC, base d'organisations non-profit de l'IRS). Ils utilisent des techniques de segmentation par machine learning pour dédoublonner la base. En gros, c'est l'équivalent d'une base Compass générée par scrapping. Leur IA permet aussi d'analyser le comportement des clients pour identifier leurs besoins via un réseau de neurones entraîné avec des données comportementales de milliers de clients. Enfin, ils engagent ensuite la conversation avec eux via des mails personnalisés et en analysant les réponses des clients pour gérer la suite des événements[432].

Alphalyr (2014, France) a créé une solution de reporting commercial qui s'appuie sur de la business intelligence et de l'IA. Elle génère un email de synthèse quotidien avec les indicateurs clés. Ils utilisent des réseaux de neurones pour l'apprentissage et la détection de signaux faibles.

Showpad (2011, Belgique, $89,5M) a développé un outil de coaching et d'outillage des commerciaux de sociétés plutôt orientées b2b.

Enfin, le spécialiste de l'équipement des centres d'appels **Genesys** est censé utiliser IBM Watson pour améliorer ses services en analysant le flot de données généré par les appels clients. C'est aussi ce que fait **Daisee** (2017, Australie, $6,7M).

Histoire de boucler la boucle, reste à vérifier, par analyse des sentiments, que cette montée en charge de l'utilisation de l'IA dans la relation clients n'aboutisse pas à un rejet de ces mêmes clients excédés par les robots, comme la sélection directe à l'arrivée (SDA) dans les centres d'appels !

Personnalisation

La personnalisation de l'expérience client est un objectif de nombreuses marques, revendeurs et annonceurs et les startups ne manquent pas d'imagination pour mettre en scène l'IA pour proposer des solutions allant dans ce sens. C'est le vaste business de la personnalisation et de la recommandation. Mais il y a du boulot car il est très difficile de disposer des informations relatives à chaque client ou prospect pour leur faire des recommandations pertinentes, surtout en termes de timing.

On vit au quotidien ces erreurs de ciblage avec d'innombrables scénarios de ratés. Nous sommes bombardés de reciblage publicitaire dès que l'on a cherché un produit d'un type donné. Ce harcèlement est risible. Il intervient bien évidemment même lorsque l'on a acquis le produit en question, ce que les autres sites et ad exchange ne savent pas forcément. C'est le cas pour les recommandations dans le voyage.

[432] Voir la vidéo Data Driven Sales: Building AI that searches, learns and sells de Anand Kulkarni, leur Chief Scientist, septembre 2015.

Vous planifiez vos vacances pour une destination donnée, vous y aller, et vous serez ensuite exposé à de la publicité sur cette destination pendant des années, même si vous changez à chaque fois de destination. Bref, le marketing prédictif ne sait pas gérer vos envies imprédictibles ni même vous faire de recommendation pertinente pour une autre destination !

Les méthodes les plus efficaces consistent à s'intégrer là où les clients expriment des besoins comme dans les moteurs de recherche, qu'il s'agisse de Google Search ou Bing, ou dans des sites de vente en ligne généralistes comme Amazon.

L'IA permet-elle d'envoyer un message personnalisé ? En théorie oui. Mais il faudrait qu'elle dispose de bien plus d'informations qu'elle n'en collecte aujourd'hui, sauf chez Google qui sait le plus de choses sur notre vie d'internaute.

Il faudrait aussi assembler des outils bien plus sophistiqués que ceux qui existent aujourd'hui. Comme pour notamment analyser les sentiments, le vocabulaire employé par le client lorsqu'il s'exprime dans les réseaux sociaux et s'y adapter de manière… manipulative par psycho-engineering. Les solutions en reviennent souvent aux techniques de micro-segmentation plutôt que de véritable personnalisation.

Les outils de recommandation de produits dans les sites de vente en ligne s'appuient sur du machine learning. Les catalogues de produits sont valorisés avec des systèmes de reconnaissances d'images similaires. Les sites web tirent parti d'outils d'optimisation du parcours utilisateur. Enfin, de nombreux outils automatisent ou accompagnent le rôle des conseillers commerciaux dans les centres d'appels entrants et sortants.

techniques de recommandation

filtrage collaboratif : basé sur les actions passées des utilisateurs

basées sur les contenus : et leurs caractéristiques

basées sur les utilisateurs : socio-démographie

recommandation sociale : sur la confiance

Emarsys (2000, Autriche, $55,3M) propose des outils de personnalisation tout azimut de la relation client multicanal, mis notamment en œuvre dans l'envoi d'emails ciblés, avec de l'A/B Testing s'appliquant notamment à l'heure idéale d'envoi d'emails promotionnels. Leurs outils permettent aussi de personnaliser la navigation dans le site web de la marque et notamment pour ce qui est des scénarios de sortie.

SouthPigalle (2015, France, $1,8M) vise aussi la personnalisation de l'expérience client. Ils ciblent notamment les entreprises du secteur du luxe. Il semble surtout s'agir d'une agence web avec sa boîte à outils intégrant la création de chatbots, des outils de segmentation de bases clients et de marketing prédictif pour automatiser la création de parcours client personnalisés.

NSP (2000, France) solution d'automatisation du marketing en mode cloud, qui sont partenaires d'Inria autour du machine learning appliqué au monde du tourisme. On est dans le cas d'une entreprise plutôt traditionnelle qui cherche à moderniser ses outils.

L'écosystème français a toujours été prolixe en startups b2b et b2c dans le secteur du ecommerce et du marketing. Il est donc normal d'y retrouver quelques startups intégrant de l'IA.

AntVoice (2011, France, $3,5M) propose une solution de recommandation prédictive pour les sites de ecommerce qui s'appuie sur du machine learning. La solution analyse la pondération de la relation entre Internautes et produits et s'appuie sur la théorie des graphes. Ils ont pivoté pour cibler le marché de la publicité en ligne.

Datapred (2014, France) propose également une solution d'analyse prédictive basée sur du machine learning. La société cible divers marchés professionnels dont celui de la distribution, en plus de la finance, de la logistique et de la santé. Elle permet par exemple de simuler des hypothèses marketing et leur impact sur une chaîne logistique de distribution en tenant compte d'un grand nombre de paramètres. Comme c'est souvent le cas, le lancement d'un projet requiert une bonne part de service et de personnalisation avant sa mise en oeuvre opérationnelle.

DataPublica / C-Radar (2011, France) est une société qui propose une solution en cloud de marketing prédictif B2B permettant de cibler les bons prospects. Elle s'appuie sur l'exploitation des données administratives et financières des entreprises issues de sources publiques, des sites web associés, des réseaux sociaux et des mentions dans les médias.

Ces données permettent alors de segmenter automatiquement les clients, de prioritisation de ces segments, le tout s'appuyant sur un apprentissage supervisé. L'approche permet par exemple de segmenter les startups d'un secteur d'activité donné (Medtech, Fintech). La société est une autre participation d'IT Translation. Elle a été acquise par l'éditeur de logiciels Sidetrade en juin 2017.

D'autres startups se positionnent sur ce créneau comme **Compellia** (2015, France), qui analyse des sources données ouvertes et identifie des événements clés de la vie des entreprises pour créer des listes de prospects qualifiés, sachant que le processus est spécifique à chaque marché.

Il y a aussi **TinyClues** (2010, France, $25,4M), une startup plus établie qui utilise des solutions de machine learning pour identifier les produits que les clients de sites de vente en ligne sont le splus succeptibles d'acheter, histoire d'optimiser les campagnes marketing ciblées au niveau du ciblage comme des messages et des offres. A Camaieu comme client.

Search'XPR (2013, France, $2M) était une startup créée à Clermont-Ferrand ayant créé le concept de "sérendipité psycho-cognitive" issu d'une thèse soutenue en 2010 par Jean-Luc Marini, l'un des cofondateurs de la société. Le concept était mis en œuvre dans la solution Oorace, destinée au commerce en ligne et même traditionnel. Elle permettait d'analyser l'état d'esprit du consommateur et d'évaluer sa réceptivité à des propositions commerciales inattendues, affichables notamment dans des offres ciblées s'apparentant à du "retargeting publicitaire" un peu moins bourrin que celui de Criteo. Le tout s'appuyait sur de l'analyse syntaxique des sites visités et du parcours du visiteur, associant algorithmes et sciences cognitives analysant les "émotions" des utilisateurs, avec à la clé une augmentation des taux d'achat et du niveau des paniers moyens. Le service était fourni sous la forme d'APIs en cloud. Reste à savoir si les algorithmes relèvent réellement de l'IA et comment ils fonctionnent. Je parle au passé car la startup a fermé boutique. Ce sont des choses qui arrivent !

Dictanova (2011, France, $1,3M) est une société nantaise à l'origine d'une solution d'analyse textuelle des feedbacks clients dans les réseaux sociaux ou sites de vente en ligne, en liaison avec les outils de CRM pour optimiser la relation client. Les techniques utilisées comprennent l'analyse sémantique de textes et la classification automatique. La solution est fournie en cloud. C'est une autre participation d'IT-Translation.

Modizy (2012, France, $275K) propose un assistant d'achats dans la mode basé sur un algorithme d'intelligence artificielle. Modizy propose aussi une place de marché reliant consommateurs et marques.

Evénementiel

L'organisation d'événements comme des conférences et des salons peut aussi faire appel à des outils exploitant l'intelligence artificielle. C'est un secteur encore naissant mais porteur d'innovations.

Il peut s'appuyer sur la capacité à capter des éléments de comportement de l'audience ou des visiteurs, pour les analyser et orienter ces derniers.

Nous avons déjà vu le cas de **Datakalab** qui analyse les émotions visibles de l'audience d'un événement, permettant d'identifier les messages et intervenants les plus impactants de manière chiffrée (on s'en rend compte souvent sans outil…).

Il y a aussi **Bziit** (France), une startup bordelaise qui veut associer l'IA et l'événementiel. Leur plateforme logicielle sert à traiter et classifier les données événementielles collectées pour détecter diverses anomalies sur le visitorat en fonction des investissements dans la promotion et le ciblage, analyser l'audience et ses réactions, son profiling avant et après l'événement, le tout avec génération de recommandations pour optimiser le rendu d'un événement.

Les données exploitées sont celles des flux physiques (visitorat par salon ou stand, durée de visites, timing), les flux dans les réseaux sociaux (discussions, hashtags, tonalité, profil des influenceurs) et les flux CRM (contacts générés, leads, etc). Ils visent les marchés du tourisme, de la distribution, des collectivités locales et de l'événementiel.

Il faudrait y ajouter une catégorie déjà bien couverte, celle des robots d'information pour les lieux publics, qui font aussi partie d'une stratégie marketing.

Ressources humaines

Peut-on injecter de l'intelligence artificielle dans les ressources humaines ? Il semble que oui, tout du moins, essentiellement dans les processus de recrutement. C'est encore en observant les créations de startups que l'on peut se faire une idée des grandes tendances du domaine, surtout aux USA, où ls entreprises n'ont peur de rien et ne se soucient pas trop d'éthique ou de valeurs humaines, malheureusement. En plus du recrutement, l'IA dans la RH peut aussi servir à gérer les talents internes de l'entreprise, à les affecter à des missions et à gérer leur mobilité interne.

Par contre, à y regarder de près, il vaudrait mieux que ces différents outils facilitent le travail des RH et des recruteurs plutôt qu'ils ne les remplacent car sinon, à ce train là, les IA auront un rôle nivelant dans le recrument et tous les originaux se feront éjecter sans compter d'éventuelles discriminations qui pourraient provenir des biais volontaires ou involontaires des données d'entraînement de ces IA. Est-ce que l'IA rend le recrutement plus humain[433] ? Pas vraiment !

Recrutements

Cela commence avec des outils d'aide à la rédaction d'annonces d'emplois efficaces et d'analyse des réponses des candidats comme chez **Textio** (2014, USA, $29,5M) (*exemple ci-dessous*). On est ici dans le domaine du traitement du langage (NLP).

[433] C'est la thèse défendue dans How AI Makes Recruiting More Human de Steven Jiang en septembre 2018.

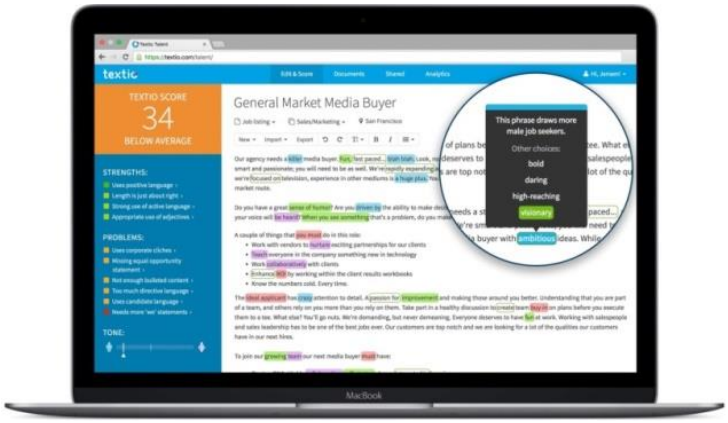
analyse automatique d'efficacité d'annonces d'emplois

+

aide à la rédaction

startup US

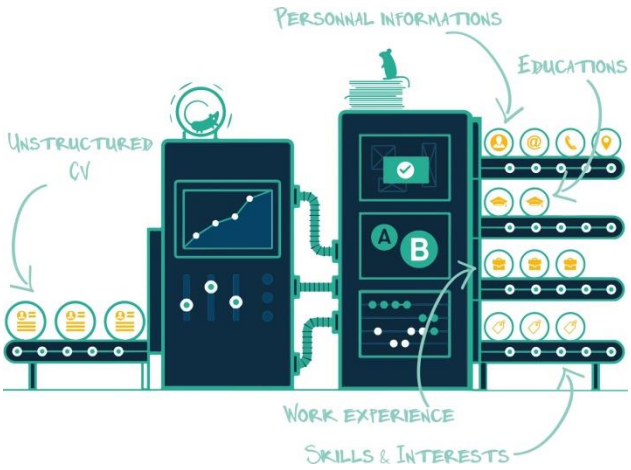
$29,5M levés



Nous avons ensuite des outils d'analyse prédictive pour identifier des talents à chasser avec **Entelo** (2011, USA, $41M) et **Gild** (2011, USA, $26M, acquis par Citadel en 2016). Ce genre d'outil s'appuie sur des techniques de prévision exploitant du machine learning et l'accès à du big data. Entelo est doté d'un moteur de recherche qui scrute les profils d'individus sur Internet pour les exploiter, à partir de 70 critères comme l'état de leur employeur (acquisition, IPO, évolution du cours de bourse, analyse de sentiment).

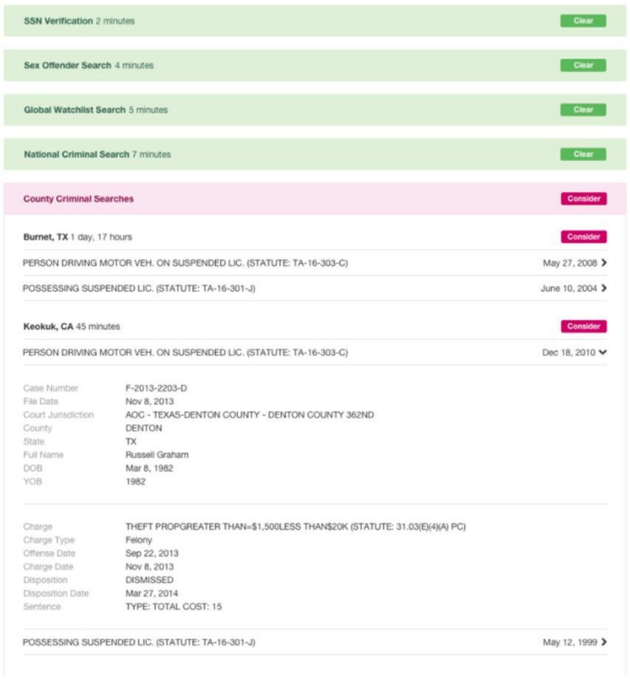
La partie IA de ce genre de solution n'est pas visible par les candidats sollicités. Ce sont des outils d'*empowerment* des recruteurs. La startup française **Clustree** (2013, France, $11,6M), lancée par Bénédicte de Raphélis Soissan, utilise aussi l'IA pour rapprocher l'offre et la demande.

Riminder (2016, France, $2,3M) est une autre startup qui propose de filtrer les CVs avec du deep learning en récupérant les informations structurées et non non structurées comprises dans les CV des candidats et dans les ressources d'Internet. Cela leur permet de faire des prévisions sur l'adéquation des CVs aux postes ouverts. Ils utilisent de l'analyse sémantique pour extraire les informations textuelles des CV et de la reconnaissance d'entités (noms d'employeurs, email, téléphone, postes occupés, formation et diplômes). Leur système est entraîné pour reconnaître et rapprocher une grande variété de fonctions.



Ils utilisent ensuite du machine learning pour faire de la segmentation automatique pour identifier les meilleurs profils.

Les propositions de candidats sont alors expliquées selon une grille de critères (expérience, soft skills, motivation, formation, etc). Les fonctions de Riminder sont aussi disponibles sous forme d'APIs pour les éditeurs de logiciels de RH. Les données d'entraînement sont issues de sources mondiales pour éviter les biais culturels locaux.

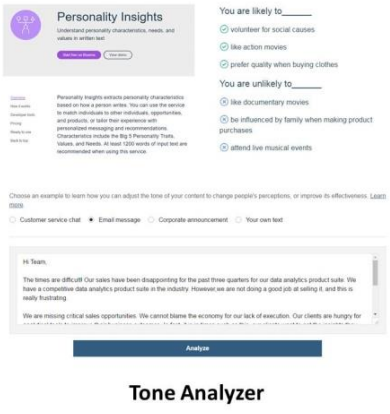


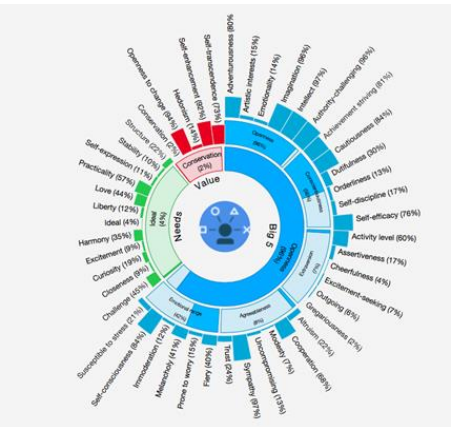
Le principe consiste à scanner toutes les sources publiques disponibles pour identifier les incohérences, bizarreries ou plus simplement, un casier judiciaire. Vu le nombre de prévenus passés par la case prison aux USA, c'est encore plus utile là-bas.

Bayes Impact (2014, France, $120K) est une société originale qui veut utiliser l'IA pour le bien public et se positionne comme une ONG. Son créateur Paul Duan s'est fait connaître en lançant un partenariat avec Pôle Emploi pour faciliter le rapprochement entre l'offre et la demande d'emplois.

IBM Watson est utilisable pour **analyser votre personnalité** à partir de vos écrits, en s'appuyant sur Personality Insights et Tone Analyzer, deux outils d'analyse de vos écrits[434] qui font partie des différentes API de Watson.

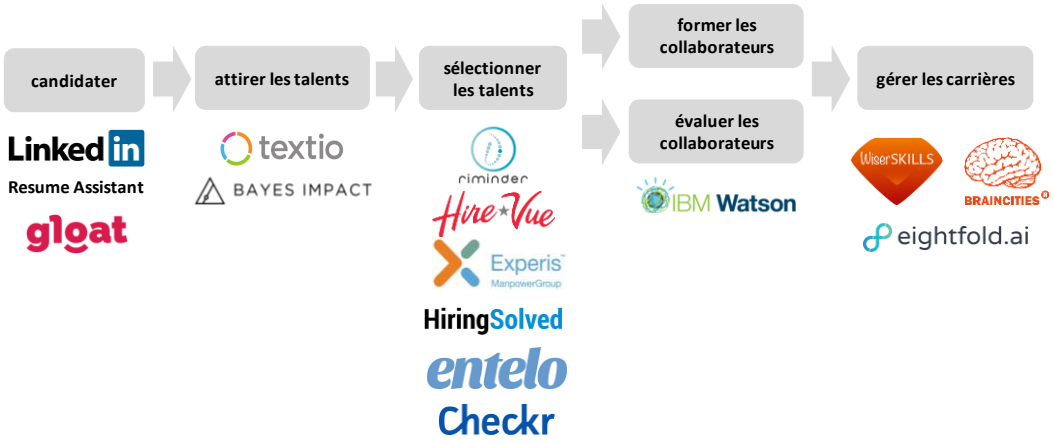
La solution permet en tout cas de détecter l'humeur de l'auteur, comme **sa tristesse**. Et peut-être d'améliorer les recrutements, tout du moins de candidats qui ont une vie publique sur Internet.





[434] Voir IBM Watson Developer Cloud, Personality Insights et IBM Watson Developer Cloud, Tone Analyzer.

L'analyse de personnalité peut aussi exploiter les vidéos d'interviews, même si le candidat parle à une machine, comme un jeu de serious gaming. C'est ce que propose l'Américain **HireView** (2004, USA, $93M). Leur logiciel analyse les visages et identifie des traits de personnalité[435]. La solution est déployée chez Unilever aux USA[436]. L'histoire pourrait se compliquer si les recruteurs se mettaient à utiliser des variations de ces systèmes d'analyse du visage comme ce prototype de Stanford qui détermine automatiquement les préférences sexuelles[437].



Le recrutement est aussi un autre terrain de jeu pour les chatbots comme celui que l'agence française **TheChatbotFactory** a déployé à la BNP.

Aux USA, **Google** est aussi présent sur ce marché avec Google Hire lancé en juillet 2017 et surtout destiné aux PME qui utilisent la G Suite, la suite bureautique en ligne de Google. Le processus de recrutement s'intègre donc naturellement dans Gmail et Google Calendar pour dispatcher les entretiens d'embauche vers les personnes disponibles dans l'entreprise. Google Hire permet aussi d'optimiser le texte de ses offres d'emploi, ne serait-ce que pour l'indexation dans Google Search ! Hire gère enfin un vivier de candidats pour les faire ressortir du lot lors de nouvelles ouvertures de postes. A vue de nez, c'est plus une application de travail collaboratif qu'une IA mais Google met en avant le fait qu'elle s'appuie sur de l'IA, ne serait-ce qu'au niveau de ses fonctions d'analyse du langage.

D'un point de vue symétrique, **Microsoft** et sa filiale **LinkedIn** utilise sa base pour conseiller les utilisateurs lors de leur rédaction de CV (vidéo) dans Word, le Resume Assistant. Ce n'est qu'un des outils que l'éditeur pourrait proposer afin de faciliter le recrutement. Et comme ils disposent de la base de LinkedIn, ils peuvent l'exploiter à fond !

Experis IT, une filiale de Manpower, a développé une solution qui permet d'avoir un entretien d'embauche avec un avatar, qui n'est pas sans rappeler les tentatives du genre réalisées par L'Oréal et le Crédit Agricole en 2007 sur Second Life. Leur Digital Room qui est un cocon d'isolation du candidat analyse son visage et sa voix ainsi que ses réponses aux questions posées. Le système va alors générer un rapport au recruteur et au candidat. C'est visiblement adapté au recrutement de fonctions d'exécution. Ca fait envie ! Cette innovation digitale était présentée lors de Viva Technology en mai 2018. L'inconvénient est de nécessiter cet engin, mais ont devrait pouvoir faire la même chose à distance avec un simple laptop.

[435] Il existe divers moyens d'analyse. Des chercheurs arrivent à le faire juste en analysant le mouvement des yeux ! Voir Eyes tell all for artificial intelligence - even your personality, new study finds, juillet 2018.

[436] Voir l'étude de cas documentée par HireVue : https://www.hirevue.com/customers/unilever-finds-top-talent-faster.

[437] Voir This AI knows whether you're gay or straight by looking at a single photo, septembre 2017.

Gestion des carrières

WiserSkills (2016, France) a conçu une solution qui sert à cartographier les compétences à développer dans l'entreprise et à exploiter cette information pour préparer les collaborateurs à s'y adapter. C'est un outil en ligne où le collaborateur décrit lui-même ses compétences et ses sources de motivations. Il ne s'agit pas seulement de renseigner ses seules aptitudes métier. L'outil était en test à la Société Générale en 2017.

Leena AI (2015, Inde) est un chatbot pour RH qui répond aux questions des salariés de l'entreprise. Il s'intègre dans Slack ou Facebook Workplace. Cela couvre surtout les processus standardisés de l'entreprise comme sur sur les congés et les notes de frais.

Braincities (2013, France) propose une « IA bienveillante » avec des applications dans les RH, la finance ainsi que pour les Smart Cities. Et notamment un outil de machine learning pour analyser les parcours de carrière et le matching entre collaborateurs et équipes dans les métiers techniques. Un autre outil de la startup permet d'analyser les échanges textuels dans l'entreprise pour les associer à ses éléments culturels. Histoire de détecter les comportements déviants ?

eightfold.ai (2016, USA, $23,75M) est un outil de gestion de pool de talents basé sur l'exploitation de données internes et externes à l'entreprise. Au bout du compte, tout cela est censé accélérer l'ensemble du processus de recrutement, et qui plus est, de réduire les erreurs de recrutement. Le produit a cependant l'air d'être, comme pour nombre de startups US, adapté au marché américain, ne serait-ce que tu fait des liens avec les logiciels de HR qui sont eux-mêmes nord-américains et pas forcément déployés en France.

Enfin, **Pôle Emploi** utilise sa base de données de 8,5 millions de demandeurs d'emplois (en 2016) pour proposer des parcours de métiers aux demandeurs d'emploi s'appuyant sur leurs compétences existantes et prédire leurs chances de retour à l'emploi[438]. Ils prédisent aussi les chances qu'un emploi d'un recruteur soit pourvu.

Comptabilité

La comptabilité et l'expertise comptable font appel aux logiciels depuis des décennies. Est-ce l'IA peut impacter ce métier ? Il est à vrai dire transformé de plusieurs manières. La première est la saisie des écritures comptables par les entreprises elles-mêmes, au gré de leur informatisation qui est devenue inévitable même pour les TPE. En gros, le travail de saisie est passé des experts-comptables vers les clients.

La seconde partie du métier relève de la vérification des comptes et des règles comptables, juridiques et fiscales. Elle est réalisée par des experts-comptables puis par des cabinets d'audit. Nombre de vérification de ces règles sont déjà automatisables par les logiciels traditionnels. L'IA peut ajouter son grain de sel en identifiant des bizarerries dans les comptes comme de recettes ou surtout des dépenses ou recettes qui sortent de la normale.

Les solutions de comptabilité servent aussi à gérer sa trésorerie, à la planifier, à créer des business plans en cas d'augmentation ou de modification de la structure de son capital. Elles permettent de préparer les négociations avec les banques en cas de demande de prêts.

Voici un petit échantillon de startups dans ce secteur sachant que je n'ai pas analysé l'offre d'acteurs étalis comme Intuit, Sage, Cegid ou EBP.

Agicap (2016, France) propose une solution de gestion comptable en cloud pour TPE qui s'appuie sur de l'IA pour gérer sa planification de trésorerie, pour savoir quand embaucher, comment gérer son encours client, ses emprunts, etc. Le tout s'appuie sur des techniques de machine learning assez classiques.

---

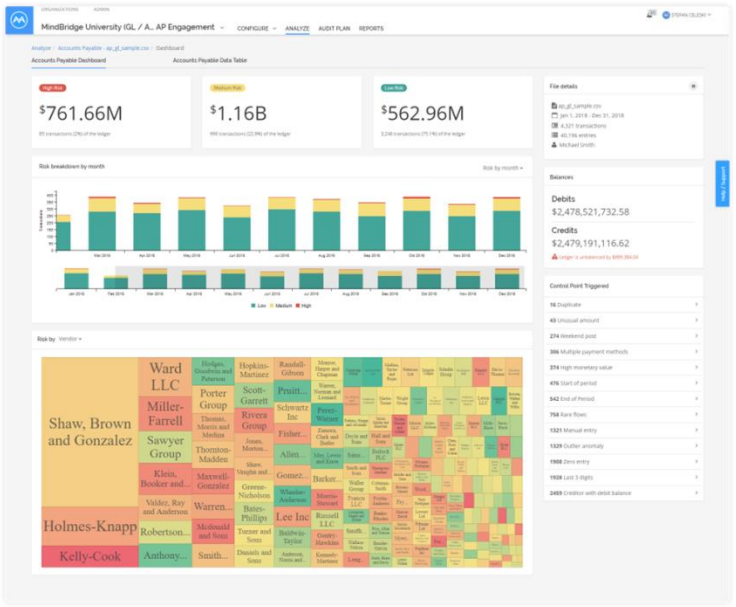
[438] Vu dans L'intelligence artificielle en entreprise du Cigref, octobre 2018 (40 pages).

iPaidThat (France) est une autre solution de comptabilité qui s'appuie sur de l'IA pour faire du rapprochement de factures et paiements dans la comptabilité. Une IA à base de machine learning, mais il n'est jamais évident de savoir ce qu'elle peut bien faire de concret par rapport aux techniques logicielles habituelles.

Dooer (2015, Suède, $4.6M) ajoute de la reconnaissance d'image, en fait de l'OCR, aux processus comptables, pour scanner et reconnaître le contenu des factures et notes de frais.

Suplari (2016, USA, $13,4M) se positionne sur un domaine voisin de la comptabilité, la gestion des fournisseurs. C'est aussi à base d'IA, non précisée. Dire qu'un tel logiciel s'appuie sur de l'IA est similaire à dire que c'est … du logiciel. Pour être un peu plus spécifique, il faudrait au minimum indiquer quelle technique est utilisée et quelles données servent à entraîner l'IA.

Mindbridge AI (2015, Canada, $12,7M) a développé Ai Auditor, une solution d'audit de comptes à base d'IA combinant machine learning, des modèles statistiques et un moteur de règles.



Une attention particulière est portée sur l'analyse de l'encours client. Le logiciel permet aussi l'interrogation des comptes en langage naturel (vidéo de démo).

Ils documentent bien cela dans la présentation How the World's First Auditor Based on Artificial Intelligence Is Driving Change in the Auditing Process, 2018 (16 slides).

White (2015, France) est une startup qui permet la saisie automatique de pièces comptables pour l'expertise comptable et l'audit. L'outil est capable de comprendre la structure du document et de le traiter convenablement dans son environnement. Il va au-delà des solutions traditionnelles d'OCR (optical characters recognition).

Dhatim (2008, France, 5M€) automatise la gestion des factures et le contrôle des déclarations sociales avec comme premiers clients les opérateurs mobiles (pour les factures), d'autres (pour les déclarations sociales) tout comme le travail de saisie dans les cabinets d'expertise comptable[439] (vidéo). La solution permet notamment d'automatiser la saisie et l'affectation comptable et d'éviter de générer des incohérences dans les déclarations sociales et les pénalités qui vont avec les contrôles qui sont eux inévitables. La solution s'appuie sur une combinaison de centaines de règles métiers et de machine learning lit les documents non structurées et déclenche des actions automatisées. La disruption est d'actualité dans l'expertise comptable !

Georges (France) est un robot expert comptable destiné aux professions libérales et pour 24€ TTC par mois (vidéo).

Des solutions existent depuis longtemps et évoluent pour intégrer des briques d'IA, en général de machine learning, pour identifier des phénomènes anormaux dans les comptes des entreprises. Nous avons de l'automatisation de comptabilité avec **Smacc Hypatos** (2015, Allemagne, $3,9M) qui cible les TPE et PME. Et puis de l'optimisation de planification financière d'entreprises avec **Anaplan** (2006, USA, $300M), **Adaptive Insights** (2003, USA, $22,5M) et **Trufa** (2013, USA, $17M).

[439] Voir Dhatim présentera Conciliator Expert, le premier logiciel de saisie comptable complètement automatisé grâce à l'Intelligence Artificielle lors du 73ème congrès de l'Ordre des Experts Comptables, octobre 2018.

Citons enfin l'Américain **H&R Block** qui a mis IBM Watson dans les mains de ses conseillers fiscaux « brick and mortar » pour optimiser la fiscalité de ses clients. C'est un modèle qui sera probablement de plus en plus courant : des AI qui améliorent la productivité des professionnels dans les services mais ceux-ci conservant un contact humain avec les clients.

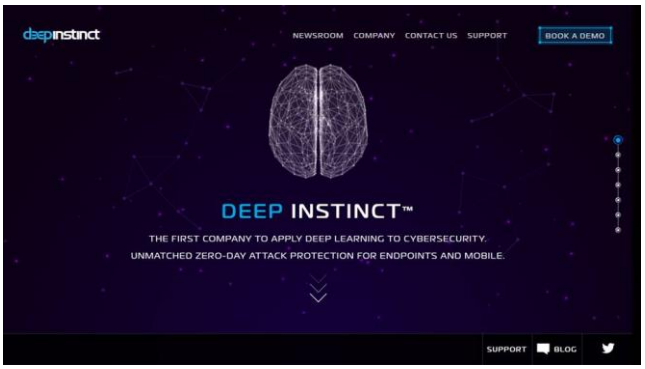
Cybersécurité

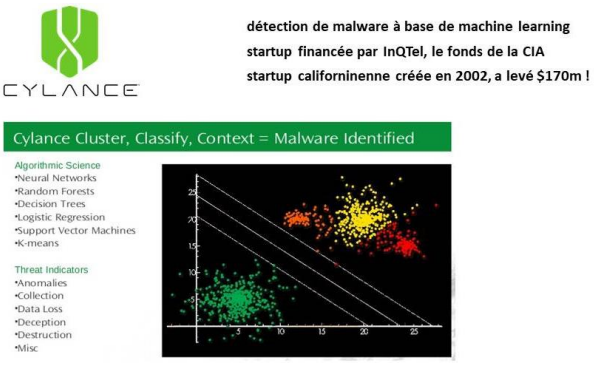
La cybersécurité est aussi un beau terrain de jeu pour l'IA, pour détecter les menaces qu'il s'agisse de spam d'email, de mail de phishing ou d'identification de vulnérabilités diverses dans les réseaux et systèmes d'information, notamment les « non malware attacks » (remote login, attaques par scripts et macros, etc). Mais elle génère aussi ses propres menaces et vulnérabilités, notamment dans les réseaux de neurones qui peuvent être trompés par des adversaires mal intentionnés.

Protections à base d'IA

C'est encore du côté des startups que nous allons faire le tour des outils de la cybersécurité exploitant peu ou prou de l'IA.

Les tentatives de phishing sont détectées par **GreatHorn** (2015, USA, $8,83M) ou avec **Lookout** (2007, USA, $282M) qui sécurise les mobiles avec un modèle prédictif. Les malwares sont détectés avec du machine learning par **Cylance** (2012, USA, $297M).





DeepInstinct (2014, USA/Israël, $37M) protège les systèmes contre les failles de sécurité récentes ("zero day threats"). Ce serait la première startup à exploiter le deep learning - avec des GPU Nvidia - tandis que la plupart utilisaient du machine learning jusqu'à présent pour faire de l'analyse multifactorielle des menaces en lieu et place de l'utilisation de bases de signatures de virus. **Interset** (2015, Canada, $24M) est dans le même créneau.

DarkTrace (2013, UK, $179,5M) utilise le machine learning pour détecter les menaces d'intrusion dans le système d'information de l'entreprise. Ils exploitent notamment du machine learning non supervisé pour détecter les anomalies de comportement à tous niveaux, dans les serveurs, desktop/laptops et réseaux, y compris les "zero day threats", ces attaques lancées immédiatement après la publication de failles dans les systèmes d'exploitation et infrastructures. Ils neutralisent ensuite les menaces avec leur "antigène" maison. L'ensemble s'interface avec les systèmes industriels au standard SCADA.

Dans le même genre, **Recorded Future** (2009, USA, $57,9M) utilise le machine learning pour détecter les menaces de sécurité en temps réel. La solution Deep Armor de **SparkCognition** (2013, USA, $56,3M) utilise aussi le deep learning et pour protéger les objets connectés dans l'industrie et les applications critiques. **CrowdStrike** (2011, USA, $481M) est une autre solution (en cloud) de protection des infrastructures d'une entreprise, détectant les attaques. **Ogo Security** (France) Senieest aussi sur ce créneau, protégeant les ressources informatiques des PME.

Des startups comme **Onfido** (2012, UK, $60,3M) vérifient l'identité de clients de service en ligne. C'est de la détection de fraude basée sur du machine learning et du prédictif.

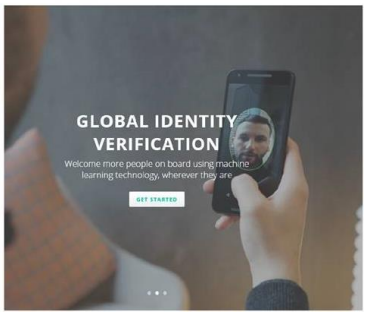
onfido

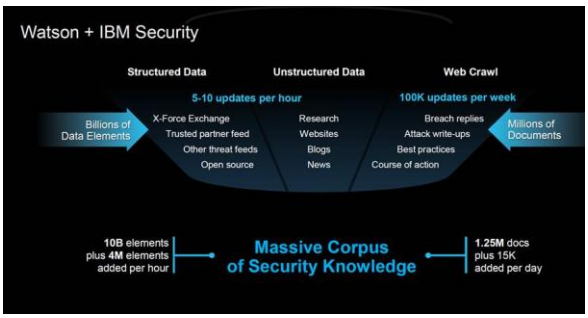
vérification d'identité multi-critères : pièces d'identité, reconnaissance du visage, géolocalisation

startup de San Francisco

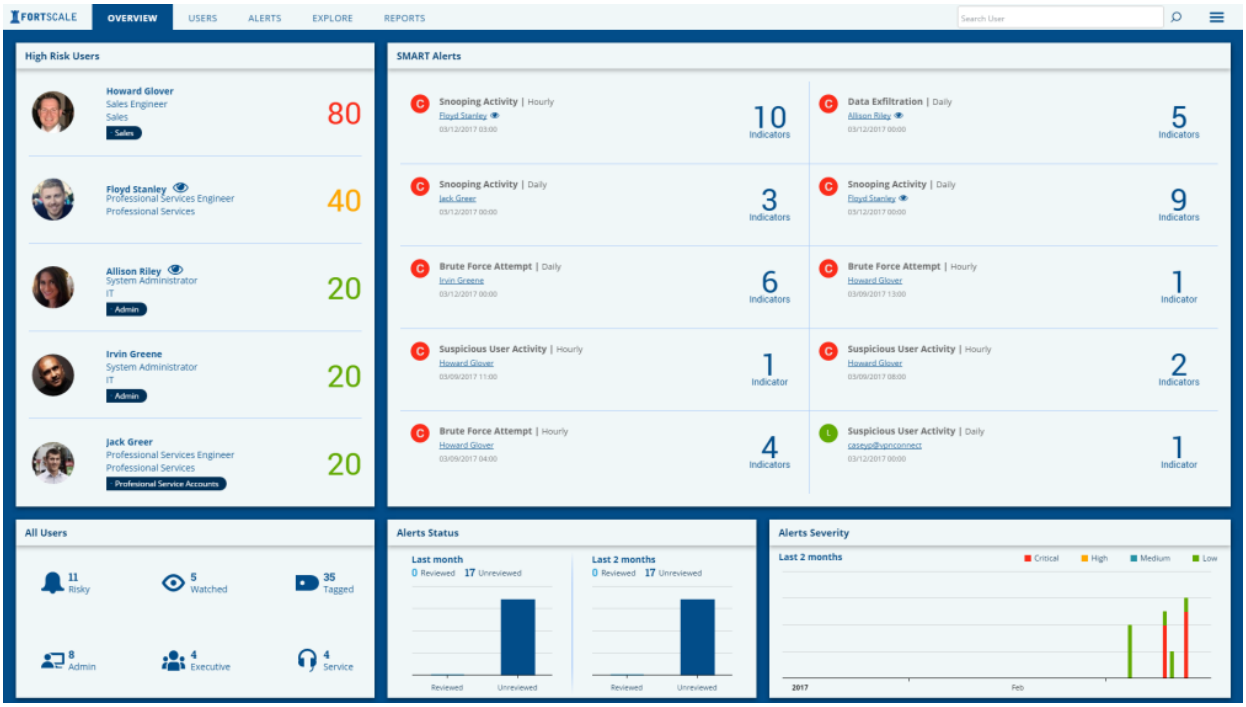
créée en 2012

$30m levés





Fortscale (2014, Israël, $23M, acquis par RSA Security en 2018) identifie de son côté les menaces internes dans les entreprises, avec sa solution User & Entity Behavioral Analytics (UEBA). Il va détecter des comportements suspects comme la copie de fichiers de grande taille sur des clés USB ! Dans les pays où ce genre de surveillance est autorisé !



IBM qui met Watson à toutes les sauces l'a aussi décliné dans la cybersécurité. Leur QRadar Advisor with Watson analyse toute la littérature disponible sur la cybersécurité pour aider les entreprises à détecter et circonscrire les menaces. La solution détecte les attaques dans les noeuds de réseaux, analyse les incidents et gère les réponses aux détections d'incidents de sécurité.

Des startups se sont aussi spécialisées sur la cybersécurité des objets connectés. C'est le cas de l'américain **SparkCognition** (2013, USA, $38,9M) dont l'offre s'articule autour de DeepArmor, une solution d'antivirus qui ne s'appuie pas sur un dictionnaire de signatures à mettre à jour régulièrement. Elle est complétée par SparkPredict qui teste de nombreux paramètres et variables de fonctionnement de systèmes embarqués pour détecter leurs failles de sécurité.

Beyond Security (1999, Israël) propose une solution dans le même registre qui teste tous les effets de bord de protocoles réseaux et logiciels pour identifier des trous dans la passoire des objets connectés. Les opérateurs télécoms sont aussi intéressés par d'autres formes de fraudes. Ainsi, Orange utilise la solution **Skymind** (2014, USA, $6,3M) pour détecter la fraude aux cartes SIM en exploitant les logs d'appels via un réseau de neurones utilisant un autoencodeur.

CybelAngel (2013, France, 4M€) a développé une solution de screening en continu du dark web sur les menaces pesant sur les entreprises clientes. La société aurait déjà Sanofi, Louis Vuitton et L'Oréal comme clients.

DrawBridge (2010, USA, $68,7M) a créé DrawBridge Identity Platform, un système d'identification des utilisateurs multi-devices à base de machine learning et connecté aux sources d'informations de l'entreprise.

IronScales (2013, Israël, $8M) utilise le machine learning pour détecter les tentatives d'intrusion par phishing.

CloudConstable (Canada) propose une solution de protection à base d'IA pour les utilisateurs vulnérables, notamment les enfants et les personnes âgées. Le tout avec un service en cloud sur abonnement.

Enfin, citons une approche originale de l'IA, la stylométrie qui permet d'identifier les développeurs de code en général, et de code malicieux en particulier ! Une nouvelle arme contre les pirates pour l'instant encore en laboratoire de recherche[440]. L'algorithme aurait une précision de 83% ce qui est un peu faible. Il permettrait, en plus de la capacité à identifier les auteurs de virus et malware, de lutter également contre le plagiat.

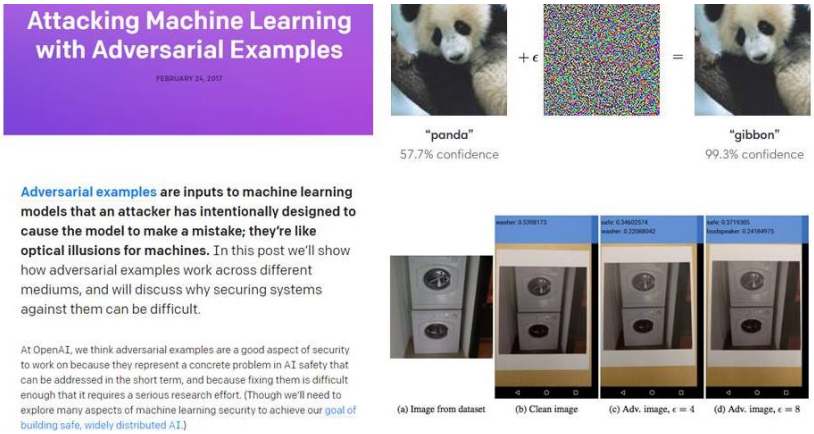
Nouvelles menaces

L'intelligence artificielle va générer de son côté de nouvelles menaces. En effet, les algorithmes de machine learning et de deep learning peuvent être retournés contre eux-mêmes par des pirates, en étant alimenté par des données bidouillées qui altèrent leurs sens. Ces attaques peuvent intervenir tout d'abord au niveau des capteurs ou des réseaux pour injecter des données modifiées.

[440] Voir De-anonymizing Programmers via Code Stylometry, avec comme contributrice Aylin Caliskan-Islam, août 2015 (17 pages) et Grâce à l'apprentissage automatique, le style des programmeurs est facilement reconnaissable, août 2018.

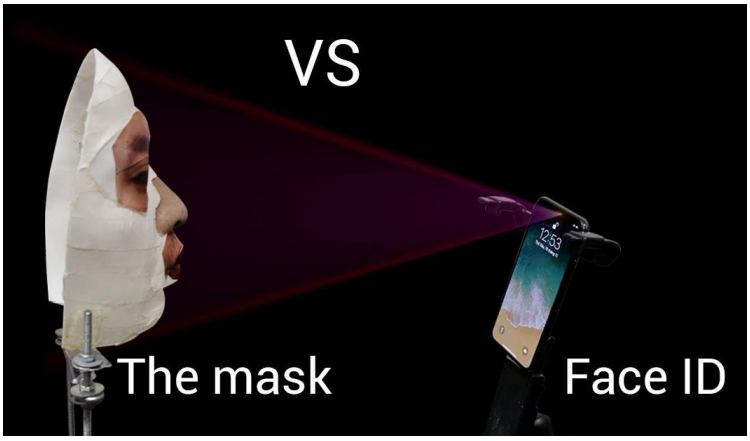
Les réseaux de neurones de vision artificielle peuvent être trompés avec des images modifiées par une technique à base de deep learning voisine de la stéganographie, qui n'en change pas l'apparence pour la vision humaine (*exemple ci-contre*)[441].

Ainsi, un panda légèrement modifié devient-il un singe pour l'algoritme de deep learning.



Cela vient des méthodes de réseaux de neurones convolutionnels et de leurs techniques de représentation hiérarchiques qui ne correspondent pas à la décomposition fonctionnelle humaine mais à des méthodes que l'on pourrait qualifier de « plus mathématiques » et qui sont contournables.

Le système de login de l'iPhone X par reconnaissance 3D du visage peut lui aussi être trompé par une fausse tête en plastique où sont collés des images du nez, de la bouche et des yeux de l'utilisateur, dans le genre Mission Impossible. Le tout, pour $150 ! Ce hack provient de la société de cybersécurité viétnamienne **Bkav**[442]. Il a été réalisé quelques jours après la mise sur le marché du nouvel iPhone X fin 2017 !



[441] Voir Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples, février 2016. Et un exemple plus récent avec Query-efficient Black-box Adversarial Examples, de Andrew Ilyas, Logan Engstrom, Anish Athalye et Jessy Lin, avril 2018 (8 pages) qui présente une technique équivalente mais très rapide. Et puis aussi High dimensional spaces, deep learning and adversarial examples de Simant Dube, avril 2018 (29 pages), qui décrit les modèles mathématiques de perturbation d'images par réseaux adversariaux. Puis Adversarial Patch, mai 2018 (6 pages) qui décrit une méthode d'attaque ciblée de parties d'images.

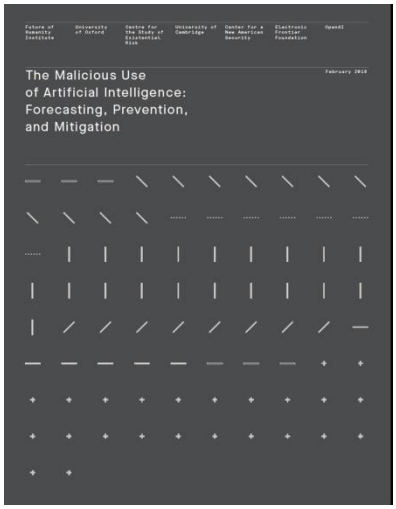
[442] Voir Bkav's new mask beats Face ID in "twin way": Severity level raised, do not use Face ID in business transactions, novembre 2017 et la vidéo associée.

Histoire d'être équitable, juste après cet exploit, c'était au tour du face login de Windows 10 d'être hacké par une simple photo[443] ! A l'envers, d'autres chercheurs ont réussi à modifier des images de visages pour qu'elles ne soient pas reconnues par des IA, mais toujours reconnues par l'œil humain[444].

Dans un autre domaine, le robot **Pepper** de Softbank Robotics serait aussi très vulnérable aux attaques de pirates du fait de nombre de failles de sécurité au niveau réseau[445].

Ces nouvelles menaces liées à des détournements de l'IA vont faire émerger à leur tour de nouvelles parades.

Des chercheurs ont déjà trouvé le moyen d'encoder des images JPEG pour éviter leur modification destinée à tromper les réseaux convolutionnels. La **Blockchain** pourrait aussi être utilisée pour garantir la chaine de confiance de transport de l'information alimentant les IA.



Bref, c'est comme dans l'armement. Des mesures de défenses amènent à la création des contre-mesures et de leurs propres contre-mesures dans une course sans fin ! Dans la cybersécurité, la tranquilité ultime est une vue de l'esprit !

Tout cela est bien documenté dans le rapport The Malicious Use of Artificial Intelligence - Forecasting, Prevention, and Mitigation qui met en avant les risques sécuritaires générés par l'IA et les moyens de commencer à les prévenir (*ci-contre*)[446].

---

[443] Voir Windows Hello Facial Recognition Bypassed with a Photo de Ryan Whitwam, décembre 2017. Mais la startup **TrueFace.ai** semble avoir trouvé une parade, aussi à base de deep learning.

[444] Voir Natural and Effective Obfuscation by Head Inpainting, 2018 (16 pages).

[445] Voir Le robot Pepper, nid à vulnérabilités de sécurité de Dominique Filippone, mai 2018 et la source de l'article : Adding Salt to Pepper - A Structured Security Assessment over a Humanoid Robot, 2018 (8 pages).

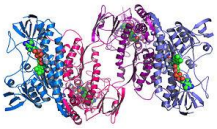
[446] A ce sujet, voir Encore une fois, les commentaires sur un rapport sur les IA passent à côté de l'essentiel, en privilégiant l'angle de l'effroi de Aymeric Poulain Maubant qui commente les réactions effarouchées et anxyongènes à ce rapport.

Applications métiers de l'IA

Dans cette grande partie, nous allons faire un tour d'horizon des usages de l'intelligence artificielle dans un bon nombre de marchés verticaux. Ils sont tous concernés de près ou de loin mais avec des rythmes d'adoption qui dépendent de nombreux paramètres comme la densité d'innovations dans leur domaine, la quantité de données qu'ils gèrent, le nombre de startups, la santé économique du secteur et sa fragmentation qui peut ralentir la propagation des innovations lorsqu'elle est trop élevée.

Sans grande surprise, les secteurs d'activité qui exploitent le plus l'IA sont ceux qui génèrent le plus de données, comme ceux de la finance, des transports, de la distribution ou de la santé. Le moins outillé est celui de l'éducation, du fait de sa fragmentation très élevée et de la difficulté à tenir les promesses un peu trop optimistes des IA qui accompagnent le parcours des élèves de manière personnalisée. Les métiers de la relation client exploitent surtout les outils de traitement du langage. Ceux de la data chiffrée font appel à du machine learning. Et les activités ayant un lien avec le monde physique utilisent beaucoup la vision artificielle et la robotique.

transports



santé

manufacturing

finance

assurance

agriculture

utilities

construction

distribution

tourisme

juridique

médias

éducation

services publics

défense et sécurité

Ici encore, l'inventaire couvrira divers projets clients et sera principalement alimenté par les offres de startups du monde entier. Elles ne sont pas toujours bien documentées, notamment d'un point de vue quantitatif, ce d'autant plus, que nombre d'annonces portent sur des « proof of concepts » et pas forcément sur des solutions déployées à grande échelle. Mais elles fournissent une bonne indication des usages et des tendances.

Transports

Le marché des transports est vaste avec le transport routier, aérien, ferroviaire, fluvial et maritime.

Tous ces secteurs sont transformés de près ou de loin par l'IA. Par exemple, sans que cela transparaisse, il est probable que les systèmes d'optimisation du transport maritime par containers soit de plus en plus optimisés par des techniques de machine learning, voire de deep learning, remplaçant des techniques traditionnelles. Le *yield management* des compagnies aériennes bénéficie aussi de ces avancées en intégrant a minima du machine learning à défaut de deep learning.

C'est surtout le transport automobile qui progresse le plus grâce à l'IA, via les véhicules à conduite assisté ou autonome que nous allons examiner de près. Ainsi que l'interconnexion entre les transports et leur optimisation pour permettre aux villes de résoudre les problèmes de congestion et de pollution. Ces évolutions seront d'ailleurs accélérées avec l'arrivée des véhicules autonomes qui n'auront plus vraiment de lignes de dessertes comme les bus pour ceux qui seront partagés.

Route

Les véhicules autonomes constituent sans conteste un défi technologique difficile à relever tout comme un facteur de changement énorme pour la vie de tous les jours et pour un grand nombre de secteurs d'activités, dans la ville intelligente, la construction, les télécommunications, les contenus et les assurances pour ne retenir que les plus évidents.

Techniquement parlant, les véhicules autonomes sont des robots capables d'atteindre un objectif en tenant compte de leur environnement et d'imprévus. A la différence des robots humanoïdes, ils sont cependant bien plus matures, même s'ils ont fort à faire pour interagir avec les Humains. La raison est simple : même si c'est une tâche complexe, faire rouler un véhicule sur une route, malgré toutes les contraintes que cela représente, présente moins de difficultés que de se mouvoir dans l'espace en 3D et d'interagir avec l'environnement physique.

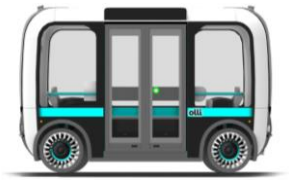
Dans une voiture, la surface de contact est relativement simple et limitée : un plan et des roues ! Plus est faible le nombre de degrés de liberté, plus l'automatisation est facile à gérer. C'est pour cela que les métros automatiques comme les lignes 1 et 14 à Paris sont déjà monnaie courante ou que les avions volent le plus souvent en pilote automatique, sauf lors des phases de décolage et d'atterrissage qui pourraient d'ailleurs être très souvent également gérées avec le pilote automatique.

Cela explique aussi pourquoi des minibus autonomes comme ceux des français **Navya** et **EasyMile** peuvent circuler dans certains lieux publics où les autres véhicules ne circulent pas[447].

L'hétérogénéité des véhicules circulant génère une complexité que les véhicules autonomes doivent gérer et en particulier lorsqu'ils doivent interagir et réagir aux passants, cyclistes et autres engins pilotés par des humains. Plus c'est hétérogène, plus c'est complexe.

Navya
Arma
autonome

EasyMile
EZ10
autonome



Local Motors Olli
imprimé en 3D, dialogue avec passagers géré avec IBM Watson IoT

D'où l'idée de mener des tests avec 100% de véhicules autonomes[448] !

---

[447] Les véhicules de **Navya** sont opérationnels dans des lieux protégés en Australie, Nouvelle Zélande, à Taiwan, Singapour, au Canada, au Luxembourg, au Danemark, en Autrice, en Suisse et à l'aéroport d'Heathrow à Londres. Aux USA, ils sont déployés à Las Vegas et doivent l'être à Disney World en Floride. En France, on en trouve à Paris, Lyon et Nantes. Les **EasyMile** EZ10 (2014, France, $22,1M) sont testés en Californie, Australie, à Taiwan, en Estonie, aux Pays-Bas, en Estonie et en Allemagne. Il faut compter aussi avec les Olli de **Local Motors** (2007, USA, $250K) ainsi qu'avec les K4 L4 Appolong du constructeur **King Long** qui sont équipées du système d'exploitation de véhicule autonome Apollo OS de **Baidu** et roulent jusqu'à 70 km/h.

[448] Cela aurait du sens de lancer un appel d'offre pour des villes de taille intermédiaire pour y expérimenter une conduite 100% autonome, avec des partenaires industriels. Il faudrait pour cela disposer aussi de véhicules utilitaires autonomes et, pourquoi pas, revoir la structure de la voirie. Il est probable que cela sera un jour expérimenté en Chine ou aux USA. En attendant, on peut compter sur l'expérience de **Transpolis** qui recréé les conditions d'usage de véhicules autonomes dans une zone dédiée à l'Est de Lyon et au Nord de l'Aéroport Saint Exupéry. Transpolis est une *joint venture* associant divers industriels dont Renault Trucks, Colas, Vibratec, Groupama et la Caisse des Dépôts. La Chine va carrément construire de toutes pièces une ville de plus de 2 millions d'habitants sur 100 km^2, équipée des dernières technologies, dont des véhicules autonomes, Xiong'an New Area, à 90 km à l'Est de Beijing.

Les véhicules autonomes sont une « réalité progressive » avec des niveaux d'autonomie étalés entre le niveau 3 (« sans les mains», le niveau 4 (« sans les yeux») et le niveau 5 (« l'esprit tranquille »). Elle existe. Elle est démontrée.

Si elle n'est pas encore courante, son contexte d'utilisation crédible s'agrandit d'année en année.

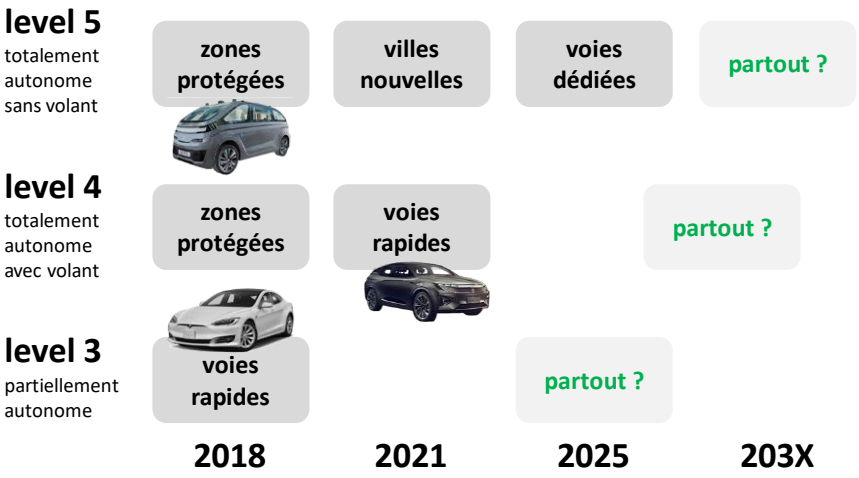
Society of Automotive Engineers' Levels of Vehicle Automation

SAE Level and Name	Definition	Steering and Acceleration	Monitoring of Driving Environment	Dynamic Driving*	System Capability
Level 0 No Automation	Full-time performance by the human driver of all aspects of dynamic driving	Human driver	Human driver	Human driver	n/a
Level 1 Driver Assistance	Execution of either acceleration or steering by a system with the expectation that the human driver will perform all remaining aspects of dynamic driving	Human driver and system	Human driver	Human driver	Some driving conditions
Level 2 Partial Automation	Execution of one or more driver assistance systems with the expectation that the human driver will perform all remaining aspects of dynamic driving	System	Human driver	Human driver	Some driving conditions
Level 3 Conditional Automation	Automated driving system performs all aspects of dynamic driving with the expectation that the human driver will respond appropriately to a request to intervene	System	System	Human driver	Some driving conditions
Level 4 High Automation	Automated driving system performs all aspects of dynamic driving under most driving conditions	System	System	System	Some driving conditions
Level 5 Full Automation	Full-time performance by an automated driving system of all aspects of dynamic driving under all roadway/environmental conditions	System	System	System	All driving conditions

*Dynamic driving consists of operational and tactical aspects of driving (changing lanes, using signals) but not the strategic aspects (waypoints/destination finding).

On passera très graduellement de l'autoroute en conduite semi-autonome à la conduite en route traditionnelle, puis en dernier ressorts, en ville. Elle méritera alors pleinement son appellation d'automobile !

La phase la plus délicate sera d'intégrer la conduite autonome dans des villes embouteillées et surtout hors des USA (Naples, Calculta, Shanghaï, la place de l'Etoile à Paris) et à les faire cohabiter avec des véhicules à conduite traditionnelle, sans compter les deux roues et les piétons. Le problème sera techniquement plus simple à gérer lorsque 100% des véhicules seront à conduite automatique dans les villes.



Pour régler les problèmes d'hétérogénéité, on interdira à long terme la conduite manuelle. C'est le stade 5 de la conduite autonome.

Les progrès de ces dernières années résultent d'efforts qui ont démarré en 2002 lors de défis lancés aux chercheurs américains par la DARPA[449]. Celui-ci consistait à faire rouler en mode autonome des véhicules sur un parcours de 240 km dans le désert du Mojave entre la Californie et le Nevada et en touchant en premier la ligne d'arrivée. Avec à la clé, un prix de $1M.

[449] L'histoire est racontée dans Inside the races that jump-started the self-driving car de Alex Davies, 2017 et The Races That Jump-Started the Self-Driving Car de Wired, 2017 (14 minutes). Voir aussi la video The great robot race (52 minutes).

Le premier défi fut organisé le 13 mars 2004. Aucun véhicule ne put faire la course en entier. Le véhicule de la Red Team de l'Université Carnegie Mellon roula juste moins de 12 km. Il fallut attendre le second défi du 9 octobre 2005 pour voir celui-ci relevé avec succès, par l'équipe de **Stanford**[450], suivie de celle de **Carnegie Mellon** associée à l'**Université de Pittsburgh** en Pennsylvanie[451]. 40 équipes avaient été sélectionnées en tout pour cette course qui faisait 212 km, également en Californie.

Un troisième défi fut organisé en 2007 avec $2M de prix. Il s'agissait de rouler en milieu urbain. Le gagnant fur l'équipe Tartan Racing associant l'Université de Carnegie Mellon University et General Motors avec Boss, une Chevrolet Tahoe modifiée[452] (slide *ci-dessus*). Le second de la course fut une équipe de Stanford utilisant une Volkswagen Passat modifiée.

Dans les contraintes à respecter, il fallait suivre le code de la route californien, être entièrement autonome, la route à suivre n'était fournie par la DARPA que 24 heures avant la course dont il fallait respecter les étapes, les véhicules ne pouvaient pas s'arrêter plus de 10 secondes, devoir pouvoir rouler sous la pluie et dans le brouillard, sans GPS, éviter les collisions de toute nature, pouvoir circuler dans un parking et faire un U-turn.

Dans les dix années qui suivirent, les progrès furent plus laborieux et incrémentaux. **Google** lançait son projet de véhicule entièrement autonome vers 2010[453], un véhicule sans volant de niveau 5. Depuis, la filiale Waymo d'Alphabet, la maison mère de Google, fait l'acquisition de véhicules autonomes pour déployer une flotte de taxis autonomes aux USA avec 20 000 SUV I-Pace chez **Jaguar** en mars 2018 et 62 000 chez **Fiat Chrysler** en mai 2018.

En 2015, **Tesla** lançait la fonction Autopilot de niveau 3 pour rester dans sa file d'autoroute. Elle s'apputait sur une caméra et un système embarqué de l'israélien Mobileye, sans LiDAR.

En 2016, Tesla**,** démontrait que ses voitures autonomes pouvaient faire un trajet complet de manière automatique, au-delà des fonctionnalités de l'Autopilot qui est surtout censé servir à rester dans sa voie sur autoroute (conduite automatique de niveau 3). Voir la vidéo 1 et la vidéo 2 avec une Tesla X. Certes, les rues empruntées ont un trafic très faible, elles sont très larges et aucun piéton n'est visible, comme souvent aux USA. Des situations que l'on rencontre plus rarement dans les villes européennes. Les démonstrations des Google Car sont du même acabit même si elles circulent plus lentement que les Tesla.

---

[450] Dans l'équipe des gagnants, on trouve notamment Joshua Anhalt, maintenant chez Uber, Hong Bae chez Faraday Future (un constructeur Chinois filiale de LeEco vu aux CES 2016 et 2017), Chistropher Baker, dans projet iSee du MIT et d'autres encore qui sont pour la plupart dans l'écosystème des véhicules autonomes.

[451] Il fallut attendre 2017 pour comprendre pourquoi l'équipe de Carnegie Mellon avait échoué et obtenu que la seconde place en 2005. Il s'agissait d'un problème de filtre dans le moteur, qui n'avait rien à voir avec la fonction de conduite autonome. Voir Carnegie Mellon Solves 12-Year-Old DARPA Grand Challenge Mystery - Highlander narrowly took second in the 2005 DARPA Grand Challenge. Just last week, CMU figured out why de Evan Ackerman, octobre 2017. De son côté, l'équipe de Cornell avaité été pénalisée par un GPS défaillant.

[452] Leur performance est documentée dans Autonomous Driving in Urban Environments: Boss and the Urban Challenge, 2008 (42 pages).

[453] Voir le TED de Sebastian Thrun Google's driverless car, 2011.

Audi commercialise depuis 2018 aussi ses A8 motorisées avec un V6, disposant de conduite semi-autonome de niveau 3 avec encore plus d'autonomie que l'Autopilot de Tesla, notamment à basse vitesse. Mais elle n'est pas encore autorisée avec ce niveau d'autonomie aux USA.

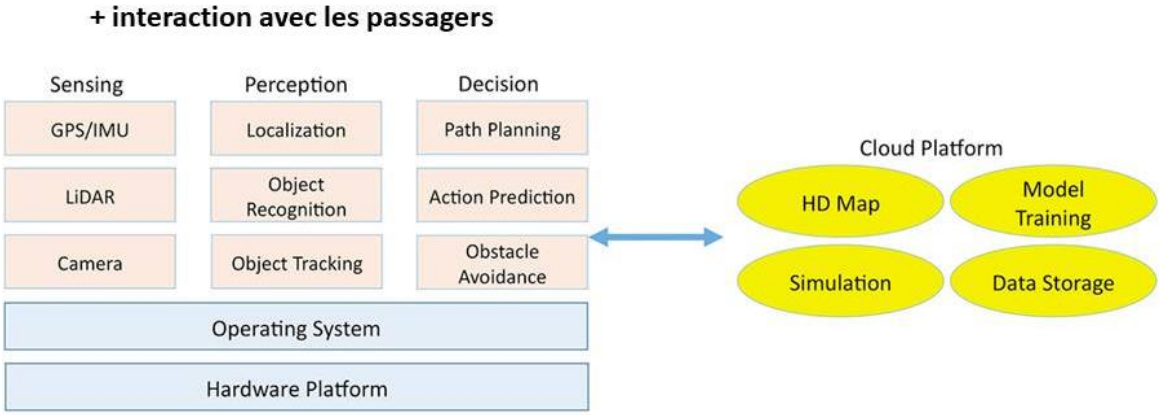
En 2015 et 2016, plusieurs expériences de conduite autonome de camions ont été réalisées en Europe, avec notamment **Volvo**. Des milliers de kilomèters ont été parcourus par une série de camions sur des voies rapides traversant plusieurs pays.

Il faut aussi creuser derrière les effets d'annonce. Ainsi, **Uber** annoncait lancer son premier service pilote de voitures autonomes à Pittsburgh en septembre 2016 avec des **Ford** Fusion. Mais les véhicules étaient tout de même pilotés, ou tout du moins contrôlés, par des conducteurs dans un premier temps ! Une expérience menée à San Francisco avec 16 véhicules de tests **Volvo** XC90 PHEV a ensuite tourné court fin 2016 après une interdiction par la municipalité de la ville. Uber a alors déplacé ses véhicules en Arizona, plus accueillant.

Les effets d'annonce ont aussi porté sur les camions autonomes. **Volvo** faisait déjà des tests routiers en 2016 en Europe. **Tesla** faisait l'annonce de son Semi en novembre 2017 (*ci-contre*) mais ils ne seront pas commercialisés avant longtemps au rythme où vont les choses chez ce constructeur. Les précommandes étaient toutefois lancées avec une petite avance de $15K. Il comprend une cabine pour le livreur !

Uber démarrait de son côté des tests de camions autonomes en mars 2018, issus de l'acquisition de la startup **Otto** (2016, USA), l'acquisition ayant eu lieu la même année que sa création, et relevant d'une d'une « acquihire »[454].

Ces tests portaient sur l'autonomie de la conduite sur longue distance, les conducteurs reprenant le contrôle des véhicules hors des autoroutes en en ville. Tout ça pour abandonner la partie en juillet 2018 et se rabattre exclusivement sur les voitures autonomes.



[454] Otto avait été cofondé par Anthony Levandowski, un ancien de l'équipe Waymo de Google chez qui il était entré en 2007 après avoir participé au projet de deux roues autonome Ghostrider de l'Université de Berkeley qui avait concourru au Grand Challenge de la DARPA en 2004 et 2005. La cofondatrice d'Otto était la Française Claire Delaunay, qui était aussi passée par Google et est maintenant VP Engineering chez Nvidia.

Un grand nombre de techniques sont mises en œuvre dans les véhicules à conduite assistée ou autonome : de nombreux capteurs (ultra-sons, radar, vidéo, LiDAR), des processeurs dédiés aux systèmes de vision artificielle temps réel (Mobileye, Nvidia, …), des télécommunications (la 5G jouera un rôle clé dans la communication entre véhicules, et entre véhicules et infrastructures connectées), des services en cloud (cartographie 2D et 3D des environnements, cartographie pour déterminer la route) et des systèmes experts de prise de décision.

Les progrès récents sont dûs aux avancées parallèles dans tous ces domaines. L'écosystème qui se met en place fait intervenir de nombreux acteurs spécialisés et créant des produits qui deviennent des plateformes comme les processeurs de Nvidia.

Pour comprendre son environnement, un véhicule autonome doit disposer d'une vision stéréoscopique ou 3D. C'est aujourd'hui le rôle des LiDAR avec leur laser tournant mais ils sont pour l'instant trop chers, coutant plusieurs milliers d'Euros l'unité. Leur marché est dominé par le californien **Velodyne** et quelques copycats chinois.

La bataille en cours consiste à créer des LiDAR dits « solid state » n'ayant pas besoin de pièces mécaniques mobiles rotatives comme les LiDAR existants. Nombre de startups comme **Quanergy** (2013, USA, $135M) ou **LeddarTech** (2007, Canada, $123M) proposent ce genre de solution, mais avec des angles de vue limités qui obligent à cumuler plusieurs capteurs pour disposer d'un angle de vue équivalent aux LiDAR rotatifs. L'autre solution consiste à utiliser des capteurs 2D traditionnels et du deep learning pour interpréter les scènces. C'est ce que fait **Mobileye** (1999, Israël, $515M) qui s'est fait gober par Intel pour $15,3B en 2017.

Les systèmes embarqués intègrent le plus souvent un GPU Nvidia adapté au deep learning d'interprétation des images générés par ces capteurs divers. La résolution des images traitées par ces systèmes est encore médiocre, ce qui limite leur précision, mais suffit aux usages actuels. Elle s'améliorera sans doute avec les progrès à venir de ces GPU et autres processeurs neuromorphiques.

Tesla est probablement le constructeur qui a le plus de véhicules semi-autonomes en circulation avec ses Model S et le mode Autopilot qui est régulièrement mis à jour. La fonction Autopilot qui a des équivalents chez d'autres constructeurs n'est pas la seule qui automatise certaines tâches de la conduite.

On peut aussi compter sur :

- Le **parking automatique**, comme avec Park4U de Valeo qui est installé sur de nombreux véhicules haut de gamme de marques allemandes et françaises. Fini les créneaux difficiles à réaliser !
- Le **maintient dans sa file** sur voie rapide (Lane Keeping Agents), une des fonctions clés de l'autopilot des Tesla S.
- Les **manœuvres** avec des agents capables de doubler un véhicule et d'autres qui permettent de sortir de la voie rapide. Il existe aussi des agents qui évitent les collisions.

Le premier accident mortel d'un véhicule équipé d'une fonction de conduite assistée est intervenu mi 2016 en Floride. Un procès s'en est suivi qui a dédouané Tesla. Le conducteur n'avait pas respecté les consignes de sécurité et les alertes[455]. Mais le camions blanc dans lequel la Tesla s'était encastré n'était pas facile à éviter pour le capteur Mobileye de la Tesla. Le constructeur a donc fait évoluer sa configuration en multiplant les capteurs, passant notamment de un à huit capteurs RGB. L'absence de LiDAR pourrait cependant rester un handicap dans ces véhicules.

D'autres accidents mortels impliquant des véhicules autonomes ont eu lieu, améliorant la courbe d'expérience des industriels. Il y a eu coup sur coup un crash de Tesla en Chine en 2016 (vidéo), un véhicule autonome **Uber** qui a renversé une dame marchant avec son vélo en Arizona en pleine nuit en mars 2018 (vidéo) et une **Tesla** Model X également en mars 2018 en Californie[456]. A ce jour, il y a donc eu quatre accidents mortels impliquant des véhicules semi-autonomes[457].

Tesla S sensors
1 RGB camera
=> 8 for 360° viewing
1 radar
12 ultra-sound
+
Nvidia K1
=> Nvidia Drive PX 2

[455] Voir The driver who died in a Tesla crash using Autopilot ignored at least 7 safety warnings de Brian Fung, juin 2017.

[456] Voir Tesla Driver Died Using Autopilot, With Hands Off Steering Wheel de Dana Hull et Tim Smith, mars 2018.

[457] C'est dans Wikipedia : https://en.wikipedia.org/wiki/List_of_self-driving_car_fatalities.

Le seul cas où le véhicule était en cause était celui de l'Uber dont le système de conduite autonome avait décidé d'ignorer l'objet qui avait bien été détecté devant le véhicule par les capteurs embarqués. Le conducteur du véhicule était distrait mais il n'est pas évident qu'il aurait évité la dame avec son vélo s'il avait été au volant, vues les conditions. Dans les autres accidents, les conducteurs n'auraient pas respecté les avertissements du véhicule.

Des accidents non mortels ont aussi été constatés. Avec par exemple un minibus du Français **Navya** percuté par un camion n'ayant pas respecté une priorité dans la zone de Fremont Street à Las Vegas en novembre 2017 ou une autre **Tesla** emboutissant un véhicule de police en Californie en mai 2018 et encore une Tela rentrant dans un camion de pompier dans l'Utah en mai 2018. A chaque accident, les commentateurs remettent en question la sécurité des véhicules autonomes et des procès sont déclenchés. Or, ce qui compte n'est pas d'avoir des accidents, mais d'en avoir moins au kilomètre parcouru qu'avec des véhicules classiques et de déclencher une courbe d'expérience permettant de faire baisser ce taux d'accident au km.

Pour l'instant, il semblerait que les véhicules autonomes soient sûrs de ce point de vue-là. Il faut toujours adopter une vue statistique de la sécurité. On devra cependant tenir compte du fait que l'on tolère moins les accidents générés par des machines que ceux qui le sont par des erreurs d'origine humaine.

Un grand débat a cours au sujet de la gestion des dilemmes par les véhicules autonomes en cas d'accident, lorsqu'il leur faudra choisir entre la mort certaine du conducteur, de ses passagers et de personnes sur la route ou entre plusieurs personnes différentes sur la route (enfants, adultes, femmes, bébés, séniors, prisonniers).

Une autre variation concerne l'autonomie des véhicules. Dans la vision « côte Ouest » des USA, les véhicules sont destinés à être totalement autonomes. Dans une vision « côte Est » ou européenne, on envisage plutôt une collaboration entre véhicules et infrastructures. Cette approche est poussée dans les concepts de smart city et notamment par les opérateurs télécoms et ceux du BTP.

Ce sont en fait des expériences de pensée assez éloignées de la réalité. Bien rares sont les conducteurs humains qui ont eu à gérer de tels dilemmes[458]. Cela conduit cependant des chercheurs à proposer l'intégration de formes d'éthique dans les algorithmes et règles de fonctionnement des systèmes à conduite autonome[459].

Au-delà de ces questions théoriques d'éthique et même s'ils se déploieront par étape, les véhicules autonomes produiront des transformations radicales de l'industrie automobile et de nombreuses industries adjacentes[460].

Tout d'abord, il est fort probable que les véhicules personnels perdront de leur attrait pour une bonne part de la population, notamment en ville. Les flottes de véhicules autonomes avec une forte densité de circulation répondront plus vite à la demande en termes de temps de réponse qu'un véhicule personnel garé dans un parking qu'il faut aller chercher. Cela pourra remettre en cause la structure du métier de constructeur automobile.

Quelques autres exemples :

- Il y aura **moins de voitures en circulation** et moins d'embouteillages dans les villes dominées par les véhicules autonomes.

---

[458] Voir Robot Cars And Fake Ethical Dilemmas de Patrick Lin, Forbes, avril 2017. Qui explique que les dilemmes éthiques évoqués ne sont que des expériences de pensée théoriques qui présentent l'intérêt de pousser la réflexion aussi loin que possible.

[459] L'approche peut consister à générer un vote social pour identifier la préférence à intégrer dans l'IA. Est-ce de l'éthique pour autant ? Pas évident ! La foule est-elle toujours intelligente ? Vous avez deux heures ! Voir A Voting-Based System for Ethical Decision Making, septembre 2017.

[460] Voir l'excellent article 73 Mind-Blowing Implications of Driverless Cars and Trucks de Geoff Nesnow qui inventorie 73 transformations liées au déploiement des véhicules autonomes.

- Les passagers pourront vaquer à d'**autres occupations** dans leur véhicule, qu'il s'agisse de travail ou de loisir. Les techniques de personnalisation numérique de l'environnement des véhicules se développeront.
- Il y aura, si tout va bien, **beaucoup moins d'accidents**[461], avec un impact sur le marché des assurances d'un côté et aussi, sur les systèmes de santé qui seront en partie désengorgés.
- L'usage de véhicules autonomes sera accompagné d'une migration à la **propulsion électrique**, avec un impact positif sur la qualité de l'air dans les villes.

- Les **parkings** pourront être plus compacts, ceux-ci ne nécessitant pas d'être accessibles par les passagers.
- Cette migration se fera au prix d'un changement des **méthodes de production d'énergie**, demandant plus de production d'énergié électrique, avec des énergies intermédiaires stockables, ce que ne permettent pas forcément les énergies renouvelables issues du soleil (photovoltaïque et éolien).Des **métiers disparaitront** comme avec la fin des diligences et du cheval de transport ! Les métiers de la conduite représenteraient plus de 4 millions d'emploi aux USA[462], soit de 1,6% à 3,5% des emplois selon les états. Les radars disparaitront aussi et on n'aura plus besoin de **permis de conduire**[463] !

Driving Jobs	**Agriculture**	**Miscellaneous Jobs**	**Justice**	**Highway Related**
Taxi drivers	Tractor drivers	Traffic reporters on the news	Traffic cops	Traffic jams
Uber & Lyft drivers	Combine operators	Sobriety checkpoint people	Traffic courts – lawyers, DA, judges	Traffic signs
Delivery (FedEx, UPS, USPS) jobs	Swather operators	Auto industry lobbyists	Driver licenses	Traffic lanes
Courier jobs	Bailer operators	Stoplight installers	Patrol cars and officers	Speed zones
Bus drivers	Sprayer operators	Pothole repair people	DUIs and drunk driving	Road stripes
Truck drivers	Horse trailer drivers	Emission testers	Sobriety checkpoints	Weigh stations
Valet jobs	Grain truck operators	Road and parking lot stripers	The boot	Mile markers
Chauffeurs and limo drivers	Automated fruit harvester operators	Night repair crews	Road rage school	Guardrails
Other Jobs	**Construction**	**Vehicle Repair**	**Crash test dummies**	**Highway Repair**
Road construction flag people	Crane operators	Roadside assistance	Road rage	Traffic cones
Drivers-Ed teachers	Road grader operators	Auto repair shops	Fender benders	Road closures
Defensive driving schools	Earth movers	Body shops	Car theft	Detours
Traffic analysts	Street sweeper operators	Tow trucks	Getting lost	Stoplights
Car licensing and registration	Backhoe operators	Glass repair	Lost cars in parking lots	Pilot cars
Drivers test people	Trencher operators	Auto locksmiths	Driving tests	Flag people
Rental car agents	Cement truck operators	Transmission repair shops	Traffic stops	Merge lanes
Crash testers	Fuel truck operators	Auto part stores	Crash test dummies	Night lights for late night road repair
Specialty Vehicles	**Car ecosystem**	**Vehicle Maintenance**	**Parking Related**	**Traffic Laws**
Forklift drivers	Auto sales – new and used	Gas stations	Parking lots	Speeding tickets
Lawnmower operators	Account managers	Car washes	Parking garages	Failing to stop at a stoplight or stop sign
Snowplow operators	Auto auctions	Oil change businesses	Parking tickets	DUIs – driving under the influence
Water truck drivers	Credit managers	Detail shops	Valet services	Reckless driving
Fire truck drivers	Loan underwriters	Tire shops	Parallel parking	Driving in the wrong direction
Water taxies	Insurance agents and sales reps	Brake shops	Parking meters	Passing in a no passing zone
Ambulance drivers	Insurance claims adjusters	Emissions testing	Charging stations	Unsafe lane changes
Trash truck drivers	Insurance call center agents	Alignment shops	Handicap parking	Driver profiling

- Il y aura moins d'accidents mortels sur la route, ce qui aura un impact indirect sur la source **d'organes pour les greffes**[464] issus de donneurs en bonne santé. Un beau paradoxe à gérer.

---

[461] La route fait environ 1,25 millions de morts par an dans le monde, concentrés en Inde et en Chine en volume (260 000 pour cette dernière), puis aux USA ! Plus que n'importe quelle guerre. Il faut y ajouter entre 20 et 50 millions de blessés par an. Voir Road Safety Facts.

[462] Dans Stick Shift : Autonomous Vehicles, Driving Jobs, and the Future of Work, 2017 (40 pages).

[463] La liste de cette page provient de 128 Things that will disappear in the driverless car era, avril 2016.

[464] Voir Plus de voitures autonomes, c'est aussi moins d'organes pour les greffes d'Ian Adams et Anne Hobson, novembre 2018.

- Les **villes** pourront être réaménagées. Les temps de trajets seront plus prédictibles et l'intermodalité plus facile à mettre en œuvre. Cela rendra la vie des banlieusards plus acceptable et aura un impact sur le marché immobilier. On aura par contre encore besoin de feux de circulation pour laisser passer les piétons !

En tout cas, les équipementiers et les constructeurs se mettent tous en branle pour se préparer à cette évolution radicale du marché. Ils sentent le vent venir et veulent aussi éviter de se faire dépasser par les acteurs issus du numérique tels que Google.

C'est le cas de **Ford** qui a lancé en 2017 sa filiale commune avec Carnegie Mellon, Argo AI, dotée d'un budget de $1B. Ce même Carnegie Mellon gagnant ou second des DARPA Grand Challenges !

Renault Nissan s'est aussi lancé, l'annonçant au CES 2017. Un test de déploiement de Zoe électrique autonome est lancé à Rouen avec l'opérateur **Transdev**[465].

PSA n'est pas en reste, ayant déjà testé en France une Peugeot autonome en 2015 et un C4 Picasso en juin 2017. Dans ce cadre, le constructeur français s'est associé à **nuTonomy** (2013, USA, $19,6M) qui développe les logiciels de pilotage également utilisés par Ford ainsi qu'avec l'**Inria** en 2018 avec l'ouverture d'un Openlab avec l'Institut PRAIRIE (PaRis Artificial Intelligence Research InstitutE).

Voici quelques autres startups du secteur, qu'il est très difficile de départager. Ils utilisent généralement les mêmes bases technologiques :

- **Optimus Ride** (2015, USA, $23,3M) une spinoff du Massachusetts Institute of Technology qui développe la partie logiciel de véhicules autonomes de niveau 4, y compris de charriots élévateurs (vidéo).
- **Netradyne** (2015, USA, $16M) est un spécialiste de deep learning appliqué la la vision des véhicules avec leur plateforme matérielle et logicielle Driveri qui s'installe sur des véhicules existants pour de la conduite assistée (vidéo).
- **Drive.ai** (2015, USA, $77M), créé par des anciens de Stanford, propose aussi une plateforme de conduite autonome à base de deep learning. Ils lancaient une expérimentation de véhicules autonomes au Texas en mai 2018[466] (vidéo) qui ont la particularité d'intégrer un panneau lumineux de chaque côté pour communiquer textuellement avec les piétons pour leur indiquer s'ils peuvent ou pas passer devant le véhicule.

- **Comma.ai** (2015, USA, $8,1M) ambitionne de proposer une sorte de SDK permettant de rendre autonomes des véhicules existants avec leur OpenPilot (vidéo). Le logiciel associé est open source. Il exploite un smartphone qui surveille le conducteur pour vérifier qu'il est bien au contrôle de son véhicule.
- **Cortica.ai** (2007, Israël, $69M) se veut le champion de l'IA qui apprend toute seule, par apprentissage non supervisé et par renforcement, avec leur "autonomous AI" et leurs 200 brevets associés. Ils interviennent notamment dans la vision artificielle appliquée à la conduite auto-

[465] Voir Rouen Normandy Autonomous Lab avec Transdev et Renault Zoé, juin 2018.

[466] Voir Drive.ai is launching an autonomous ride-hailing network in Texas de Megan Rose Dickey, mai 2018.

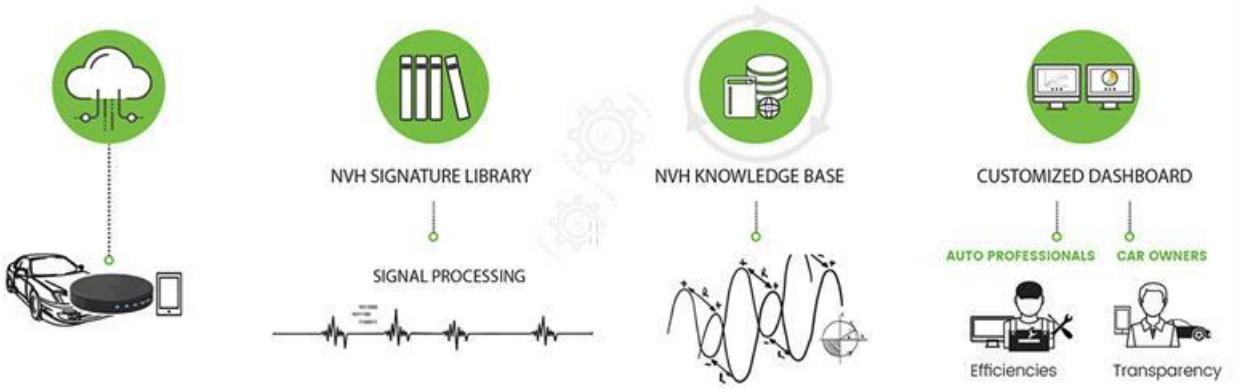
nome et notamment pour anticiper les mouvements des objets et personnes sur la route (vidéo). Leur IA est aussi adaptée au scanning de bagages et à l'exploitation de vidéosurveillance (vidéo). Mais il est bien difficile de faire la part des choses entre leur discours et les démonstrations qui semblent exploiter des réseaux de neurones convolutifs qui ont l'air d'être assez classiques pour la détection et la classification d'objets.

- **Ween** (2014, France, $1,8M) a développé une IA qui détecte les mouvements des utilisateurs dans leurs usages quotidiens et dans leurs trajets avec une solution à basse consommation. Après un apprentissage par renforcement, cette solution prévoit en temps réel l'arrivée des utilisateurs à leur prochaine destination (domicile, voiture, bureau, transport public...), ce qui permet de proposer un accueil personnalisé (mise en température du logement, éclairage bureaux, etc).

En février 2018, **Huawei** et **Baidu** collaboraient pour permettre à un smartphone Mate Pro 10 de piloter un véhicule autonome. L'idée était d'utiliser la caméra du smartphone et le NPU du chipset Kirin 970 (processeur pour le deep learning) pour reconnaître l'environnement du véhicule et éviter les obstacles. C'était testé sur une **Porsche** Panamera modifiée. Le tout s'appuyait côté logiciels sur la plateforme HiAI de Huawei (SDK mobile pour le Kirin 970), PaddlePaddle de Baidu (SDK de deep learning) et Baidu Brain (vidéo).

Enfin, citons **Apple** qui est enfin sorti du bois en 2018 pour expérimenter relativement discrètement un logiciel de véhicule autonome dans une cinquantaine de véhicule dont des **Volkswagen** ainsi que des SUV RX450h de **Lexus**.

L'IA peut aussi apporter des services en sortant du cadre des véhicules autonomes. En voici quelques exemples. Elle peut aussi servir à optimiser les trajets, notamment pour les professionnels. C'est un des domaines où **Uber** travaille, pour optimiser le temps de travail de ses conducteurs. **IBM** propose une solution « Watson on Wheels » qui optimise aussi les trajets en fonction d'informations sur la qualité de la voirie. Une analyse des données issues des smartphones, de l'enregistreur de bord, de caméras, de la vitesse et du régime moteur permet d'évaluer le comportement du conducteur et, éventuellement, de moduler ses primes d'assurance en conséquence.



L'analyse du bruit du moteur permet de faire de la maintenance préventive. C'est une fonction proposée par la startup **Otosense** (2014, USA, $1,2M), qui est basée à Cambridge (USA) et créée par le Français Sébastien Christian[467], chez **Carfit** (2016, USA, $2,3M), également créé par un Français, Henri-Nicolas Olivier (ex Modelabs et Inventel), et qui exploite l'analyse des vibrations du véhicule (schéma *ci-dessus*) ainsi que chez **Uptake Technologies** (2014, USA) qui l'a notamment expérimenté dans l'armée de terre US sur des chars Bradley M2A3[468].

[467] La startup proposait à l'origine un système transformant les alertes sonores en alertes visuelles pour les malentendants. C'est ensuite qu'elle s'est mise à cibler le marché industriel. Ils sont en test chez PSA. Leur projet AudioHound est une tablette conçue pour le technicien du garage pour réaliser des diagnostics pas analyse du bruit du véhicule.

[468] Uptake Technologies est moins spécialisée que Carfit et Otosense. Ils ciblent tout un tas de marchés : l'agriculture, le rail, la construction, l'énergie et l'industrie.

Rail

Le rail utilise aussi des véhicules autonomes et depuis longtemps, ne serait-ce que dans les métros des aéroports ou les lignes 1 et 14 du métro à Paris. La sécurité est assurée par les doubles portes. Les utilisateurs savent aussi généralement s'adapter aux automatismes et jouent moins avec les portes pour rentrer au dernier moment dans les wagons.

L'automatisation du rail pourra un jour également touche les lignes de train. A ce titre, la **SNCF,** l'IRT **Railenium**, **Alstom**, **Altran**, **Ansaldo**, **Apsys**, **Bombardier**, **Bosch**, **Spirops** et **Thales** (ça fait du monde !) ont annoncé la création de deux consortiums visant à développer d'ici 2023 deux démonstrateurs de trains autonomes, un TER et pour le transport de fret. On en est pour l'instant au stade des avant-projets. Cela entraînera entre autres une refonde des systèmes de signalisation, de contrôle-commande et l'optimisation de l'exploitation ferroviaire.

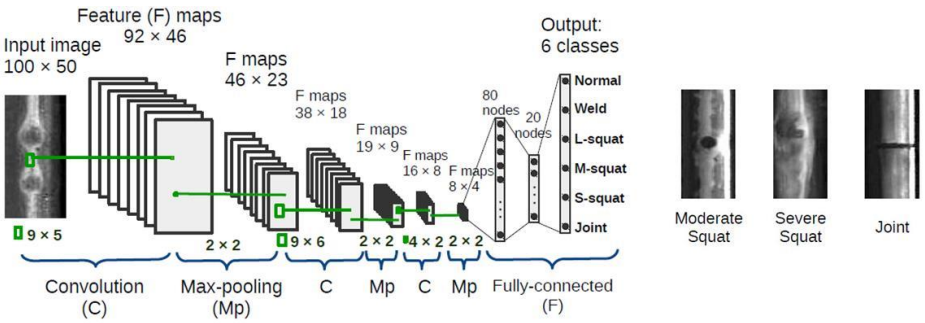
Sinon, le monde du rail est depuis longtemps un utilisateur de l'IA. Il a commencé à expérimenter les systèmes experts dans les années 1980. Il est surtout focalisé sur l'optimisation des ressources et la maintenance prédictive.

Les systèmes experts et la logique floue peut servir à la planification, à l'évaluation des retards et à la reprogrammation de trains après des retards. D'une manière générale, les transports nécessitent des outils de plus en plus sophistiquées pour gérer le matériel et les Hommes qui les opèrent.

La recherche de défauts sur les rails, les caténaires, dans la signalisation et dans le matériel roulant fait partie des domaines de prédilection de l'IA. Des accidents comme celui de Brétigny en 2013 ont accentué les efforts de la SNCF en France dans la surveillance des infrastructures. La SNCF inspecte les caténaires de ses lignes avec des drones et de la classification des images des isolateurs en céramique ou en verre.

J'ai trouvé des traces d'inspection visuelle des voies ferrées à base de réseaux de neurones à partir de 2013, ce qui correspond au début de l'engouement pour la technique[469]. Elle peut porter sur l'inspection des rails, du ballast et des environs. L'objectif est d'identifier des obstacles sur les voies, les défauts de structures sur les rails et la croissance de la végétation[470]. L'inspection peut être réalisée par des caméras installées sur des motrices dédiées à la maintenance des voies et même sur des motrices tirant les trains habituels[471].

Aux Pays-Bas, la détection de défauts sur des rails s'appuie sur des réseaux de neurones convolutionnels classiques exploitant des images à basse résolution (100x50 pixels) pour classifier une demi-douzaine de défauts répertoriés[472].



[469] Voir Automated Visual Inspection of Railroad Tracks de Esther Resendiz, John Hart et Narendra Ahuja, 2013 (10 pages).

[470] Voir Automating Condition Monitoring of Vegetation on Railway Trackbeds and Embankments de Roger Nyberg, 2015 (301 pages).

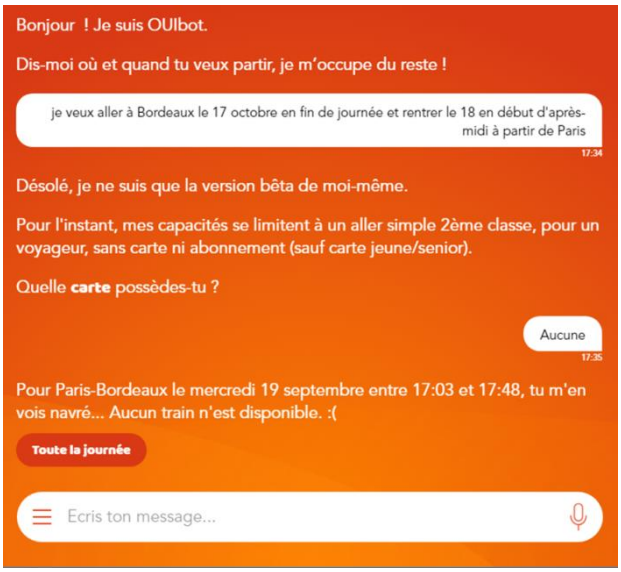
[471] Voir Digital Twin for the Railway Network Making Trains "Look" for Track Defects de Dattaraj Jagdish Rao, GE Transportation, 2018 (16 slides).

[472] Source : Big Data in railway infrastructure, 2018 (52 slides).

Mais la maintenance prédictive peut aussi exploiter de nombreuses données issues des capteurs intégrés dans le matériel roulant[473].

C'est ce que fait la **SNCF**, avec **Quantmetry** (2010, France), **Global Sensing Technologies** (2011, France) et **IBM Watson** pour la maintenance des Transiliens depuis un pilote lancé fin 2015.

Enfin, la relation client grand public passe évidemment par les chatbots, même si ceux-ci sont encore loin d'être au point comme l'illustre l'exemple ci-contre avec celui de **Ouisncf** qui ne sait pas gérer une demande sur un aller/retour et ne mémorise pas la date de la première demande[474] !



Du côté de l'offre utilisateurs, la notion d'intermodalité entraîne des évolutions de l'offre associant différents opérateurs de transports. On en profite déjà largement via les applications transverses mobiles, à commencer par Google Maps.

Geo4cast (2012, France) propose dans ce domaine des services d'optimisation des transports pour tous les modes. La startup récupère des données de déplacement issues de smartphones, est connecté avec des données automobiles, transports en commun et permet d'analyser et d'optimiser les transports entre eux pour réduire les investissements et dépenses superflues. Cela pourrait servir à terme la coordination automatique de réseaux de véhicules autonomes. Le tout exploite plusieurs technologies comme le nettoyage de données, la reconnaissance de modes de transports basée sur du machine learning, des algorithmes d'optimisation et du revenue management.

Aviation

Le monde du transport aérien est aussi un grand consommateur d'IA en pièces détachées, que ce soit pour le *yield management* appelé aussi le *revenue management* qui exploite un mix de big date et d'IA aussi bien dans les compagnies aériennes que chez Amadeus et son service de réservation en marque blanche, la gestion des opérations, la relation client et la maintenance des avions[475].

Il en va de même pour les avions eux-mêmes qui font appel à divers systèmes embarqués qui comprennent très probablement de nombreuses briques d'IA, ne serait-ce que pour la navigation.

Le marché de l'IA dans le secteur de l'aviation passerait de $112,3M en 2017 à $2,2B en 2025, avec une croissance annuelle de 46,65% probablement équivalente à celle d'autres secteur d'activité[476].

Le transport aérien génère des tombereaux de données, aussi bien au niveau des capteurs intégrés dans les avions, les données météo et d'opérations qu'au niveau des passagers transportés[477].

---

[473] Voir Using IOT to Advance Railway Predictive Maintenance de Hitachi Vantara, 2018 (28 pages) et IoT and predictive maintenance in Railways, Thalès, 2018 (14 slides).

[474] Ce qui n'a pas empêché ce chatbot de gagner le prix du « Best Robot Experience » en avril 2018. Voir aussi Comment Oui.sncf propulse son bot sur Google Home, Alexa ou Messenger d'Antoine Crochet-Damais, mai 2018.

[475] Voir Impact of AI on the Aviation Industry de Houman Goudarzi, 2017.

[476] La source est Artificial Intelligence in Aviation Market by Offering qui est citée dans la présentation Artificial Intelligence and Machine Learning in Aviation 2018 (337 slides).

[477] Voir Machine Learning in Aviation, 2013 (54 slides).

J'ai trouvé diverses applications de l'IA spécialisées dans le transport aérien comme pour prédire les retards des avions[478], pour analyser a posteriori les vols et leur sécurité au niveau de leur enveloppe de vol[479] ou pour inspecter visuellement les carlingues d'avions[480], les systèmes de reconnaissance des visages intégrés dans les bornes de l'immigration aux USA comme en France avec Paraphe, puis le **BagBot** qui remplit automatiquement les containers de valises dans quelques aéroports européens depuis 2014 et le **Skywash** qui lave les avions de toute taille depuis 1997, notamment à Frankfort en Allemagne.

DigiBot BagBot (2014) Schiphol

Skywash (1997) Putzmeister

Chez Air France, l'IA est testée pour aider les passagers à trouver les zones de checkin et de livraison des bagages dans les aéroports avec une application de réalité augmentée pour mobile, pour faire de la maintenance prédictive de sa flotte de Boeing et d'Airbus au niveau moteur et avion qui permet de remplacer des pièces critiques avant leur défaillance[481], divers chatbots pour la relation client et l'application Cargo Repair Case pour la réallocation optimisée de charges cargo dans les avions de passagers en fonction des contraintes et aléas d'exploitation.

[478] Voir Application of Machine Learning Algorithms to Predict Flight Arrival Delays, 2017 (6 pages).

[479] Voir Modeling airline crew activity to improve flight safety analysis de Nicolas Maille, 2017 (12 pages).

[480] Voir Aircraft Fuselage Defect Detection Using Deep Neural Networks, 2017 (5 pages).

[481] Voir PROGNOS - Predictive Aircraft Maintenance et la vidéo associée.

Quelques startups utilisent l'IA dans le transport aérien, telles que :

- **OpenJaw** (2002, Irlande) qui créé des outils de relation client pour les compagnies aériennes. T-Social est un chatbot développé avec les briques de traitement du langage d'IBM Watson. Il comprend un système d'escalade des cas non résolus vers des opérateurs humains.
- **Appintheair** (2011, USA, $300K) propose une application mobile d'intégration de l'expérience du voyageur aérien.
- **Yieldin** (2013, France) fait du « revenue management » à base d'IA, non précisée, probablement avec une grosse dose de machine learning.
- **Fieldbox.ai** (2011, France) aide Aéroports de Paris à réaliser le tri des bagages en analysant leurs flux.
- **PureStrategy** collecte le feedback des passagers sur leur expérience clients avec sa Automated Neural Intelligence Engine.
- **Jeppesen**, une filiale de Boeing, a créé Crew Rostering, une solution logicielle qui sert à gérer le planning du personnel navigant en intégrant différentes contraintes comme leurs préférences, leur niveau de fatigue. Elle utilise du machine learning.
- **SynapseMX** (2015, USA) permet d'optimiser la maintenance des avions avec une application mobile.
- **Mezi** (2015, USA, $11,8M) a développé un agent conversationnel qui aide les voyageurs à gérer … leur voyage top to bottom, et via SMS.
- **CrowdVision** (2009, UK) aide les aéroports à gérer leurs flux de passagers, à optimiser les temps de check-in et de passages aux points de sécurité. C'est un domaine d'application de la vidéo surveillance.

Bâtiments et Travaux Publics

Le marché du BTP comprend de nombreuses sous-catégories avec celui du gros-œuvre et du sous-œuvre dans la construction de bâtiments d'habitation ou dans le tertiaire, celui de la promotion immobilière puis celui du génie civil pour les gros ouvrages d'art ainsi que pour la construction de routes.

En France, cette activité représente plus de 170 Md€ de CA et plus de 1,4 millions d'emplois répartis dans environ 530 000 sociétés allant de l'artisan aux grands acteurs tels que Vinci et ses filiales Eurovia et BATEG, le groupe Bouygues (Colas, Bouygues Construction, Bouygues Immobilier), Spie Batignolles, Eiffage, Nexity et Quartus.

Les usages de l'IA couvrent tous ces métiers mais avec une grande variation selon la taille des structures. J'ai observé une loi empirique selon laquelle les innovations perfusent moins rapidement dans les marchés b2b très fragmentés. Les premiers industriels du secteur à adopter les nouvelles technologies sont plutôt les plus gros d'entre eux. Les artisans et TPE/PME n'ont pas les mêmes capacités d'adaptation, sauf lorsqu'il s'agit d'utiliser des outils numériques génériques (télécoms, ordinateurs, smartphones).

Construction

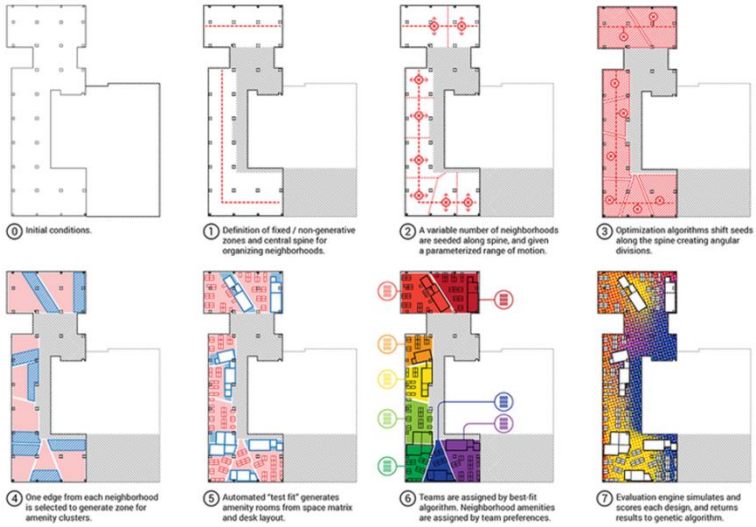
Le secteur de la construction n'a pas attendu la mode du deep learning de ces 10 dernières années pour se mettre à l'IA. Il était déjà un assez gros consommateur de systèmes experts depuis les années 1980, mais avec plus ou moins de bonheur.

En effet, nombre d'expériences de cette époque n'ont pas forcément porté leurs fruits. Comme dans de nombreux domaines, je suis donc allé à la pêche aux informations en suivant plusieurs fils d'Ariane, aidés par mon ami Google Search.

Dans Artificial Intelligence in Civil Engineering, 2012 (22 pages), des chercheurs de l'université Zhejiang de Hangzhou évoquent de nombreux cas de modélisations complexes faisant appel à des systèmes experts et de la logique floue. La logique floue est utilisée pour la modélisation d'expertise informelle. On y dénombre des applications de prévision de la solidité des routes, de l'optimisation de trajets, la prévision du comportement d'ouvrages de génie civil et de la maintenance préventive. Des outils de planification intégrent les risques et incertitudes dans les chantiers. Les solutions font appel à des algorithmes évolutionnaires, génétiques et des réseaux multi-agents.

Application of Artificial Intelligence in Construction Management (14 pages) évoque aussi la planification de constructions à base de systèmes experts.

Des équipes d'**Autodesk** présentent des outils de design génératifs permettant de définir l'organisation des bureaux. Ce projet MaRS Generative Design vient de chez Autodesk, vu en détail dans The Promise of Generative Design, 2017, et détaillé dans The Rise of the AI: Impact of AI and Machine Learning in Construction (12 pages).



Le deep learning a ensuite fait son apparition dans des applications très spécifiques. Rule-Based Fuzzy Polynomial Neural Networks in Modeling Software Process Data, 2003 (11 pages) évoque une application en Chine de prévision de la solidité de structures en béton qui utilise un réseau FPNN pour Fuzzy Polynomial Neural Networks. Dans Application of Artificial Neural Networks in Civil Engineering, 2014 (7 pages), ce sont des modèles de prévision de résistance du béton au feu par réseau de neurones.

Plus récemment, Artificial Intelligence as a Tool in Civil Engineering – A Review, 2017 (4 pages) fait un inventaire de scénarios d'usages de l'IA, mais sans mettre en avant d'études de cas précises.

On y trouve évoqué **Neuro-modex**, un modèle permettant d'optimiser les décisions de construction en fonction du lieu, de l'environnement, des conditions d'emplois, des caractéristiques du projet et de ses risques, de outils de planification avancée et des outils d'optimisation financière de la réfection de la chaussée.

Encore plus récemment, Comparison of Artificial Neural Networks and Fuzzy Logic Approachs for Crack Detection in a Beam Like Structure, 2018 (17 pages) présente une solution de détection de failles structurelles dans les ouvrages de génie civil. Elle n'a visiblement pas été employée sur l'autoroute de Gênes en Italie ! C'est à base de logique floue et de réseaux de neurones simples.

On peut citer le cas de **Doxel** (USA, $4,5M) qui utilise un petit robot chenille et des drones équipés d'un LiDAR et de caméras pour inspecter les chantiers de construction autant en intérieur qu'en extérieur (vidéo). L'outil logiciel évalue la progression des travaux et est censé détecter les malfaçons. La startup affiche pour l'instant trois clients : Kaiser Permanente (San Diego), Lucas Museum (Musée de Georges Lucas en Californie) et Sutter Health (hôpitaux, également en Californie).

Le secteur de la construction est friand de robots et autres exosquelettes pour à la fois accélérer les chantiers et réduire la pénibilité du travail tout comme celle des accidents de personnes[482]. Plusieurs startups et industriels se sont donc lancées dans la création de robots divers allant de l'automate

[482] Voir Robotics in Construction, 2016 (87 pages).

téléguidé au robot véritablement autonome, bien plus rare[483]. On assiste à un phénomène voisin de celui que nous évoquerons plus loin dans l'agriculture, à savoir, l'émergence de nombreux prototypes qui ne sont souvent pas industrialisés. Soit ils ne sont pas au point, soit leur équation économique n'est pas encore satisfaisante, soit les deux à la fois.

Les barrières à l'adoption des robots dans ce secteur d'activité sont leurs coûts d'acquisition, d'installation, d'opérations et de maintenance, la fragmentation du marché et des tâches à robotiser qui génère un besoin d'une grande variété de robots pour automatiser des chantiers[484], les difficultés de leur mise au point, l'incompatibilité avec les pratiques usuelles dans les chantiers et en dernier ressort, l'acceptation des nouveaux outils par les ouvriers[485].

Avant la construction, il a la démolition ! On trouve des robots de démolition qui sont en fait des machines téléguidés chez **Brokk** avec son modèle 180 (*ci-dessous* à gauche) et **Husqvarna** et son DXR (au milieu) et, plus intrigant, le **ERO** conçu en 2013 par un étudiant en Suède et prototypé par le fabricant de compresseurs **Atlas Copco**. Ce robot utilise un jet d'eau sous pression pour détruire les structures de murs en béton et récupérer les débris pour les recycler. Ces débrits sont aspirés et servent ensuite de matériau de construction comme gravier dans la composition de bétons spéciaux. Il ne semble cependant pas que ce robot ait été produit en série. On n'en trouve pas de vidéo le montrant en état opérationnel. C'est un cas classique de communication difficile à interpréter ! Les images présentées dans la communication d'ERO et d'Altas Copco ne sont que des modèles de synthèse en 3D ! Dommage.

Nous avons déjà évoqué dans la rubrique des exosquelettes les cas de **Ekso Bionics** et du français **DB3D** qui sont aussi présents dans le marché de la construction. Ce dernier fournit des produits expérimentés par Colas en France pour la construction de routes et la réfection des chaussées[486].

Built Robotics (2016, USA, $15M) développe des robots tracteurs dédiés à l'excavation de terrains et qui exploitent les mêmes capteurs que les voitures autonomes. Le fondateur est un ancien ingénieur de Google qui n'y a pas travaillé plus d'un an en 2010. Le principe de fonctionnement est voisin de celui d'une tondeuse de gazon autonome : l'opérateur délimite les contours du terrain à déblayer et la zone où déposer la terre, et le robot se débrouille tout seul ensuite (vidéo).

[483] Voir 8 Startups Building Robotic Construction Workers, novembre 2017.

[484] Robots de démolition, de terrassement, de levage, de soudure, de pose de ciment, de parage de surfaces, de peinture, de construction de routes, etc.

[485] Voir An investigation into the barriers to the implementation of automation and robotics technologies in the construction industry, de Rohana Mahbub, 2008 (303 pages) qui étudie les obstacles à la robotisation de la construction en Australie, au Japon et en Malaisie. L'étude a 10 ans mais ses conclusions semblent être toujours d'actualité.

[486] Voir Colas lance le déploiement des exosquelettes sur ses chantiers, mars 2018.

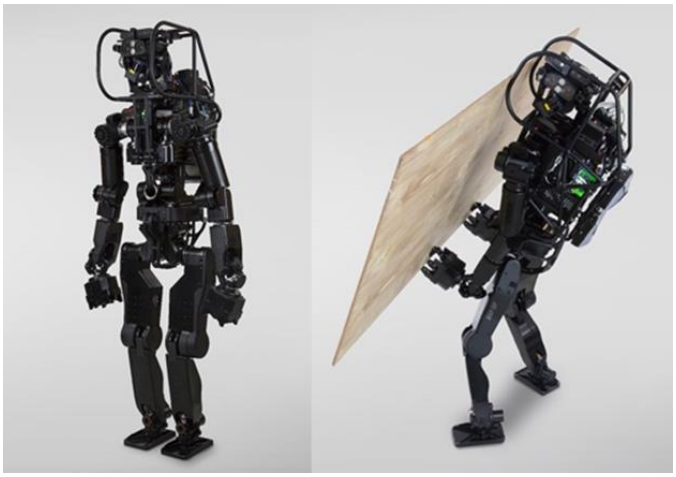
Fastbrick Robotics (2015, Australie, $10M) a créé le robot Hadrian X qui dépose des briques ou parpaings pour construire des bâtiments. Il peut ainsi construire un bâtiment de un à deux niveaux en quelques jours (vidéo). Caterpillar a investi $2M dans cette société et l'Arabie Saoudite a signé un MOU pour construire 50 000 maisons avec l'enfin d'ici 2022. Petit détail observable : le robot ne dépose pas de mortier sur les briques pendant leur pose. Ca fait un peu désordre ! Avec l'attribution de la citoyenneté à Sophia par le pays, cela fait la paire !

Construction Robotics (2008, USA, $75K) développe aussi un robot de pose de briques qui lui, sait poser le mortier autour. Le SAM100 (Semi-Automated Mason) fait gagner 80% du temps de travail du maçon et ne le remplace donc pas entièrement. C'est un « cobot » qui peut déposer 2000 briques par jour contre 400 pour un maçon. Petit détail à noter : il ne fonctionne que pour des murs droits. Il n'aime pas encore les angles !

Ce qui est gênant pour de la construction d'habitations et le rendra plus utile pour la construction de bâtiments industriels de grande taille (vidéo).

Cazza (2016, USA, $2M) a créé le X1, un robot-imprimante 3D comme il en existe déjà quelques-uns sur le marché. La machine est mobile grâce à ses chenilles. Il imprime des structures verticales ou coule des dalles horizontales. Où est l'IA dans un tel robot ? Il doit normalement disposer de capteurs et de caméras pour s'orienter. Au minimum doit-il donc disposer d'outils de reconnaissance d'images pour se déplacer (vidéo).

Endless Robotics (2015, Inde, $100K) a créé le robot WALT qui peut peindre les murs 30 fois plus rapidement qu'un peintre doué, mais a tout de même besoin de supervision humaine. Et comme nombre de robots cités ici, c'est encore un prototype (vidéo). Et puis, dans la construction, il y a aussi les robots expérimentaux de **Schaft**[487], qui fait partie du groupe **Softbank Robotics**, et le HRP-5P Humanoid Robot du laboratoire japonais AIST (vidéo). Il est capable de poser des planches. Il mesure 1,82 m et pèse 100 kg[488].



[487] Le robot de Schaft avait gagné le DARPA Robotic Challenge en 2014 (vidéo). Dans une autre vidéo datant également de 2014, on le voit réaliser diverses tâches de chantier en pièces détachées (vidéo).

[488] Voir Humanoid robot prototype HRP-5P capable of the same hard work as human being developed, septembre 2018. Le robot est encore « mono-tâche » pour l'instant.

Pour terminer sur l'usage de l'IA dans les chantiers, on pourrait intégrer la dimension humaine. Le recrutement d'ouvriers de chantiers peut dans une certaine mesure exploiter les outils évoqués dans la partie RH. La vidéo-surveillance permet de vérifier le respect des règles de sécurité. Certaines solutions permettent par exemple de s'assurer du respect de l'accès aux zones sécurisées. Les drones d'inspection peuvent aussi surveiller les chantiers et leur avancement.

Immobilier

Le marché de l'immobilier est aussi potentiellement friand d'IA, qu'il s'agisse d'améliorer les étapes de la création, de la planification des offres et de leur commercialisation. C'est un secteur d'activité où les grands promoteurs s'activent mais où sévissent pas mal de startups, en particulier pour rapprocher l'offre de la demande de logements aussi bien à l'acquisition qu'à la location. En voici quelques-unes.

Skyline (Israël, $21M) fait du benchmarking de biens immobiliers multicritères pour les investisseurs, pour leur permettre de dénicher les bonnes affaires. Les sources d'informations utilisées sont nombreuses (130) avec plus de 10 000 attributs par actif.

Enodo (2016, USA, $2,5M) propose un outil équivalent au precedent pour évaluer les investissements immobiliers prometteurs en termes de rendement.

REX Real Estate (2014, USA, $25,5M) utilise une forme d'apprentissage par renforcement pour améliorer l'efficacité du ciblage publicitaire en ligne d'annonces immobilières. Il analyse les paramètres communs de segmentation des premiers utilisateurs qui cliquent sur l'annonce pour renforcer la visibilité de l'annonce sur le segment qu'ils représentent.

OJO Labs (2015, USA, $26,5M) a créé un chatbot de relation client pour les agences immobilières. Il permet aux utilisateurs de trouver des biens qui correspondent à leurs critères de choix. Au passage, leur système reconnaît les caractéristiques des biens via les photos qui l'accompagnent.

Appartenir (France) est un service recherche immobilière qui envoie des alertes sur des annonces qui répondent à ses critères de recherche. L'IA utilisée par la startup n'est pas précisée car on n'en a pas forcément besoin pour faire des recherches multi-critères dans une base d'annonces, qui plus est lorsque sa taille est relativement réduite. Les annonces sont scrappées sur divers sites qui les publient.

First.io (2014, USA, $7,35M) propose une solution logicielle de prévision des déménagements des foyers, pour leur proposer des offres au bon moment. La société suit 700 paramètres sur 214 millions de personnes aux USA, à la fois des paramètres personnels comme ceux qui sont liés aux modifications de l'environnement (modifications de l'amménagement des quartiers, ouvertures/fermetures d'écoles, etc). C'est finalement une sorte de Cambridge Analytica adapté à l'immobilier !

CompareAgences (2012, France) intermédie la relation entre agents immobiliers et particulier dans le cadre de la vente de biens. La startup emploie 12 personnes et génère 200 000 visiteurs uniques par mois. 1000 agences immobilières sont intégrées en France. Le tout est à base de machine learning, sans plus de précisions.

Smart city

Le concept de Smart City a été créé il y a une dizaine d'années et promu notamment par IBM dans le cadre de son initiative Smart Planet (en 2008). De nombreux fournisseurs de technologie ont embrayé le pas, comme Cisco puis Nvidia. Le marché du secteur public et des villes étant important pour eux, le concept de Smart City était et reste un moyen d'adopter un discours commercial global pour approcher les collectivités locales. Diverses villes ont lancé des initiatives de ce genre, telles que Barcelone (avec Cisco), Nice (encore avec Cisco) et Amsterdam, généralement axées sur les transports. Un tas de grandes villes du monde ont emboité le pas, clamant haut et fort leurs plans de Smart City. Le *Smart City washing* est devenu un outil politique puissant.

Les diverses solutions de Smart City portent sur l'organisation des moyens de transports, de l'éclairage, de la fourniture d'énergie[489] et eau, de la sécurité, voire des systèmes de santé. Elles portent aussi sur la maintenance préventive des infrastructures. A chaque besoin sa solution et ses sources de données. Le concept de Smart City reste cependant flou car son approche systémique reste assez théorique et difficile à mettre en œuvre[490]. En effet, sauf à être lancé dans une dictature, la plupart des grandes villes connaissent un développement généralement non coordonné et l'existant est difficile à moderniser d'un coup de baguette magique. Ce d'autant plus que la Smart City est souvent une grande dévoreuse de données issues des infrastructures et des activités des citadins avec des conséquences sur la vie privée qui sont encore mal évaluées.

Des projets symboliques peuvent être lancés, tels que le **Toronto Waterfront** qui est conduit par Sidewalk Labs, une filiale d'Alphabet créée en 2015[491]. Il s'appuiera notamment sur l'usage de véhicules autonomes et de l'habitat intelligent, mais sur une zone de surface limitée à quelques hectares, en bordure du lac Ontario qui sera reconstruite entièrement (*ci-dessous*).

Le projet en est pour l'instant à l'étape de l'étude de faisabilité et de la consultation publique des besoins et idées[492] (vidéo). Si ce projet aboutit d'ici une dizaine d'années, il permettra de créer un « living lab » des idées constitutives de la Smart City et d'en conduire des évaluations objectives. En particulier pour ce qui est de l'usage de véhicules autonomes qui sera probablement plus sûr dans un environnement où ils cohabitent peu avec des véhicules conduits traditionnellement.

En pratique, la « ville intelligente » est rarement une ville intégrée. C'est une ville qui met en œuvre des services utilisant des technologies disparates pour améliorer la vie de ses habitants, réduire le coût de ses infrastructures et son empreinte énergétique. D'un point de vue pratique, la Smart City s'appuie sur des briques technologiques communes que sont les capteurs et les objets connectés, les infrastructures de télécommunications pouvant comprendre des réseaux M2M (Sigfox, LoRA) et des applications construites autour de « big data » et de machine learning pour exploiter les données

[489] Voir par exemple le projet français AMOEBA de l'IRIT (Institut de Recherche en Informatique de Toulouse), qui vise à trouver les corrélations entre prévisions de consommation d'énergie et la production d'énergie, décrit dans Use Cases of Pervasive Artificial Intelligence for Smart Cities Challenges (52 slides) et Use Cases of Pervasive Artificial Intelligence for Smart Cities Challenges (7 pages).

[490] Voir par exemple City Brain, a New Architecture of Smart City Based on the Internet Brain 2017 (8 pages), un document chinois qui cherche à créer une analogie conceptuelle entre le fonctionnement d'une ville intelligente et le cerveau.

[491] A smarter smart city - An ambitious project by Alphabet subsidiary Sidewalk Labs could reshape how we live, work, and play in urban neighborhoods de Elizabeth Woyke , février 2018.

[492] Voir la présentation Roundtable 3 qui décrit assez bien l'étendue du projet, août 2018 (118 slides).

générées par les capteurs. Le machine learning intervient naturellement pour faire des analyses de données, des segmentations, des détections d'anomalies et des prévisions.

Les applications à base d'IA qui reviennent le plus souvent dans les villes intelligentes concernent l'**optimisation des transports**, souvent routiers, principalement pour les décongestionner et en assurer la sécurité. L'étape ultime de la transformation sera le passage aux véhicules autonomes mais ce n'est pas encore à l'ordre du jour dans la majorité des projets.

La vidéo-surveillance est de plus en plus utilisée pour l'optimisation du trafic routier[493] et bénéficie des nombreuses avancées dans la reconnaissance d'image via le deep learning. L'exploitation des images de vidéo-surveillance permet d'évaluer le traffic, la vitesse des véhicules, les embouteillages, la détection d'accidents, et dans certains cas, le pistage dans la durée de véhicules suivis par la police. On peut aussi exploiter les GPS des mobiles pour prévoir les mouvements de foule[494].

Nvidia organise même depuis 2017 un « AI City Challenge » qui est focalisé sur la vidéo-surveillance, ce qui s'explique par leur offre qui associe des chipsets pour serveurs (Volta V100) et pour équiper les caméras, les Jetson TX1. Des dizaines de projets sont sélectionnés et financés qui font avancer graduellement les techniques d'analyse d'images vidéo.

RapidFlow Technologies (2015, USA) développe des systèmes de signalisation (feux) exploitant des caméras pour adjuster leur programmation dynamiquement en function du traffic et de manière coordonnée entre feux. Le système est déployé à Pittburgh, réduisant les temps d'attente de 40% d'après la startup.

Xaqt (2015, USA) exploite les données issues des capteurs intégrés dans les routes, les caméras et les données météo pour prédire l'apparition de nids de poule dans la chaussée avec une précision de 85% ! Cela permet de déclencher les réparations plus rapidement. Ou pas, selon les habitudes des services de la voirie !

Datategy (2016, France) est une startup qui lutte contre la fraude dans les transports et les parkings. Elle récupère toutes les données imaginables pour identifier les zones où la fraude est la plus forte et y envoyer les contrôleurs avec leurs terminaux de verbalisation. Cela utilise force machine learning. Dans la même veine **Vimoc** (2012, USA, $2,4M) optimise la gestion des parkings.

L'IA est aussi employée sur les autoroutes chez Vinci comme chez son concurrent SANEF.

Chez **Vinci**, le système Cyclope analyse les flux vidéo provenant de caméras existantes et réalise de la classification automatique des véhicules pour valider les tarifs de péage en cas d'ambiguïté et détecte les anomalies de comportement sur autoroute comme les véhicules arrêtés, les piétons ou les mauvais sens de circulation[495]. Le dispositif a déjà été expérimenté sur les autoroutes du Sud de la France (Escota).

Le réseau **SANEF** (autoroutes du Nord et de l'Est de la France) s'apprête à moderniser ses péages avec l'analyse des images de caméras et de LiDAR pour classifier les véhicules, tout en lisant les plaques d'immatriculations. C'est couplé à de l'analyse de trafic servant à repérer des anomalies et à remonter des alertes. Le système va entrer en production début 2019 dans deux grandes gares de péage de l'autoroute A4[496].

---

[493] Vu dans Video Analytics for AI City Smart Transportation de Ming-Ching Chang, 2017 (41 slides).

[494] C'est l'objet d'un projet de Microsoft Research publié en janvier 2017 : Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction (7 pages).

[495] Voir Vinci Autoroutes teste de premières applications métiers nourries à l'IA sur son réseau, février 2018.

[496] Voir SANEF : Test du 1er péage free flow avec lecture des plaques en France, de Thibaut Emme, Leblogauto, septembre 2018.

Utilities

Les *utilities* sont les fournisseurs d'électricité, de gaz et d'eau ainsi que les gestionnaires des déchets. Ils sont de grands utilisateurs potentiels de collecte de données et d'exploitation de l'IA pour optimiser leurs ressources de production et leurs infrastructures. Leur objectif est d'adapter dynamiquement la production à la consommation ou de modérer cette dernière si besoin est.

Il s'agit aussi de réduire leur empreinte énergétique et de réaliser de la maintenance prédictive de leurs infrastructures de production et de transport. Enfin, accessoirement, il s'agit aussi d'améliorer le service rendu aux clients. En général, le bon service se manifeste quand tout fonctionne sans que l'on ait à s'en soucier.

Les utilities exploitent l'IA dans différents autres domaines. Chez **EDF**, c'est la comptabilité des 24 000 fournisseurs et de la communication par emails avec eux qui est accélérée grâce à un assistant qui prépare les réponses à environ 60% des demandes. Chez **Engie** Home Services, qui gère l'entretien et le dépannage des systèmes de chauffage et de climatisation du grand public, la startup Vekia utilise l'IA pour améliorer sa gestion de stocks de pièces détachées[497].

DirectEnergy a créé un chatbot permettant aux particuliers de suivre leur consommation énergétique[498]. **IBM** promeut l'usage de ses données météo issues de l'acquisition de The Weather Channel pour conseiller les investisseurs dans les énergies renouvelables.

Maintenance prédictive

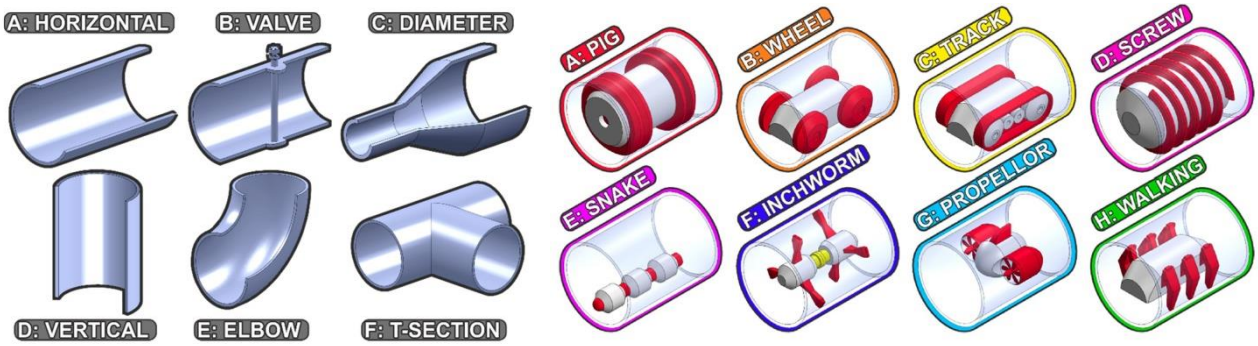
C'est le domaine où la créativité dans l'usage de l'IA semble être le plus grand et elle s'adapte aux nombreuses situations rencontrées par les utilities.

Les lignes à haute tension peuvent être inspectées par des drones qui sont pilotés automatiquement en fonction de parcours préétablis et dont les images sont aussi exploitées plus ou moins automatiquement pour détecter des défauts dans les lignes et les pylones.

Flod (2015, France) analyse leur bruit avec cymbalum-IoT pour faire de la maintenance prédictive, une solution dont les tests avaient démarré en 2016 chez Enedis.

General Electric (USA) est l'un des nombreux industriels qui proposent des engins et services d'inspection des pipelines, comme l'échographe UltraScan Duo. Le contrôle non destructif est souvent réalisé avec des sondes dénommées « smart pigs » pour "Pipeline Inspection Gauges") qui parcourent les pipelines en les inspectant de l'intérieur.

Mais on peut faire appel à d'autres engins pour les parties non linéaires des pipelines comme évoqué dans Advances in the Inspection of Unpiggable Pipelines, 2017 (13 pages).



[497] Voir ENGIE adopte l'intelligence artificielle pour sa gestion de stocks, mars 2018.

[498] Retour d'expérience: Comment Direct Energie a développé son Chatbot? (4 mn). Il n'est pas certain que l'on ait un besoin fréquent de ce genre d'outil.

Les techniques d'inspection sont variées avec de l'échographie, de la magnétographie ou de la mécanique de variation de géométrie et de diamètre[499]. Elles génèrent presque toutes des images en quantité qui peuvent être analysées par des réseaux de neurones convolutionnels pour identifier différents types de défauts.

Côté production, **General Electric** fait de la détection d'anomalies dans les turbines à gaz de centrales thermiques via la création de profils de températures en sortie de turbine. 12 paramètres sont mesurés qui génèrent une rosace de paramètres et des patterns associés détectés par un réseau de neurones[500].

L'IA peut aussi être utilisée dans les bassins hydrauliques, pour identifier en avance les ouvrages qui sont les plus fragiles face aux intempéries, comme vu dans Using artificial intelligence to locate risky dams de Sarah Fecht, août 2018.

Table 1. Handcrafted Features

ID	Feature	Description
1	DWATT	Raw turbine load
2	TNH	Raw turbine speed
3	MAX	Max TCs
4	MEN	Mean TCs
5	STD	Standard deviation of TCs
6	MED	Median of TCs
7	DIF	# diff b/w positive & negative TCs
8	ZR	Zero crossing
9	KR	kurtosis
10	SK	skewness
11	M3S	Max of 3-pt sum
12	M3M	Max of 3-pt median

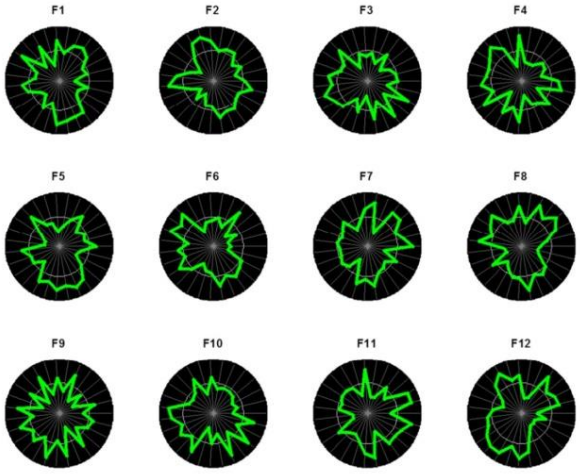


Figure 4. The 12 learned features

Engie a développé un chatbot SMS et Skype destinée à ses techniciens de maintenance des parcs éoliens, solaires et hydrauliques. Il a été développé par la société de services Eryem sur le Bot Framework et les Cognitive Services de Microsoft puis déployé dans le cloud sur Azure.

C'est un moyen différent de naviguer dans une application de visualisation de données d'un parc d'éoliennes, solaire ou hydraulique, ou des caractéristiques d'une installation donnée (sa production, la vitesse du vent, les pramètres de fonctionnement).

EDF lançait sa propre startup, Metroscope, bref, une filiale, ciblant les entreprises et proposant une solution de détection d'anomalies dans les processus industriels et de maintenance prédictive. Elle a été développée en mode agile par des équipes internes d'EDF.

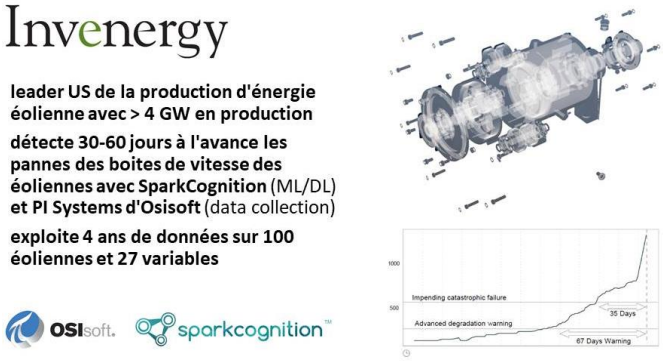
Voici quelques autres cas de figure, rencontrés chez diverses startups.

PlutoShift (2016, USA, $2,1M) utilise l'IA pour gérer les opérations des centres d'épuration d'eau. Le tout à base de machine learning. Mais c'est un éditeur de logiciel généraliste qui couvre divers besoins dans l'industrie.

[499] Neuf méthodes d'inspection en tout, inventoriées dans Critical Review of Pipeline Scale Measurement Technologies (14 pages). Voir à quoi cela ressemble dans l'édition de début 2018 dy Pipeline Technology Journal et dans Combined Crack and Metal Loss Detection Tool using Phased Array Technology, 2007 (5 pages). Evaluation of technologies to assess the condition of pipe coating on Line 5, 2018 (127 pages) évoque de son côté l'inspection des pipelines sous-marins, qui requièrent une inspection extérieure en environnement hostile sur les fonds marins.

[500] Vu dans On Accurate and reliable anomaly detection for has turbine combustors : a deep learning approach, de Weizhong Yan et Lijie Yet, 2015 (8 pages).

L'opérateur américain d'éolienne **Invenergy** exploite la solution logicielle SparkPredict issue de l'éditeur de logiciels **SparkCognition** (2013, USA, $63,2M) à base de deep learning pour déterminer les périodes de maintenance préventive des éoliennes, exploitant quatre années de données sur 100 éoliennes et 27 variables de fonctionnement. Cela rappelle qu'il faut des données de qualité pour faire du prédictif [501]!



L'application réduit le coût moyen de chaque panne en le faisant passer de $350K à $50K et la planification des réparations intervient à plus de 2 mois d'avance au lieu de 2 à 3 jours. Fin 2017, SparkCognition a été aussi sélectionné par le Département de l'Energie US pour faire de la maintenance prédictive du même genre dans les centrales au charbon[502].

Le solaire peut aussi être de la partie. **RayCatch** (2015, Israël, $7,3M) utilise ainsi le machine learning et l'analyse des données de production d'électricité pour surveiller et optimiser les installations de panneaux solaires photovoltaïques.

Gestion du grid

Les grands utilities gèrent des réseaux complexes avec diverses ressources et de nombreux points de distribution.

Les énergéticiens doivent souvent gérer l'hétérogénéité des sources de production d'énergies renouvelables (faiblement stockables) et fossiles (stockables par défaut), anticiper la production d'énergies renouvelables en fonction de la météo et des besoins de consommation pour planifier la production d'énergies d'origine fossile, piloter à grande échelle des infrastructures de production, de distribution et de consommation. Les « energy grids » peuvent aussi gérer les échanges entre réseaux de production. La notion de grid intègre aussi la supervision de la maintenance préventive, exploitant les briques d'IA vues dans la partie précédente. De grands acteurs et intégrateurs ont leur propre solution de gestion de grid, comme chez **GE** et **IBM**. Mais là encore, de nombreuses startups cherchent à rentrer dans ce marché, dont les exemples suivants.

Deepki (2014, France, $2,1M) gère la consommation d'énergie de parcs immobiliers avec 200 000 bâtiments déjà observés pour 130 clients dont une part de la Ville de Paris et une autre d'Intermarché avec 3200 magasins et entrepôts. Au bout du compte, une économie de 5% à 10% de la facture énergétique. Les outils comprennent Collect pour consolider les données d'exploitation et Ready qui les analyse et détecte notamment les anomalies de consommation d'énergie.

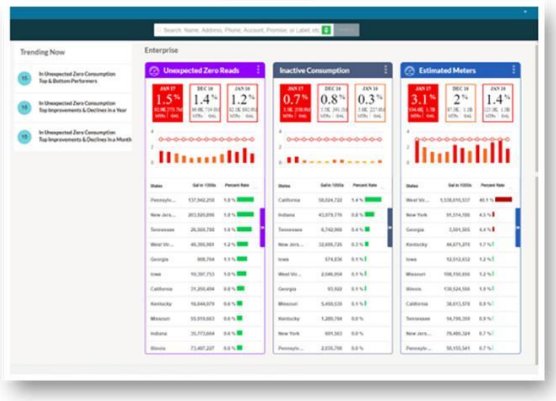
Drift (2015, USA, $7,5M) est un Enron des temps modernes qui utilise l'IA pour faire de l'échange d'énergie sur les marchés ouverts en mode peer-to-peer pour s'adapter en temps réel à l'offre et à la demande, le tout avec du trading à haute fréquence qui est justifié pour la distribution d'énergies fortement intermittentes comme l'éolien. Le service qui a démarré dans la ville de New York intègre un réseau de 3000 producteurs d'énergies renouvelables hydraulique et solaire. Le machine learning est utilisé essentiellement pour faire du prédictif, en intégrant les données climatiques et les historiques de consommation dans la zone déservie. Les clients sont les consommateurs d'énergie, aussi bien dans le logement individuel que dans les entreprises, et le prix du service est fixe.

---

[501] Voir aussi AI Used to Predict Failure of Wind Turbine Facility utilization rate improves from 21% to 23%, 2018.

[502] Le fondateur de la société Amir Hussain a publié en 2018 The sentient machine, un ouvrage de vulgarisation de plus qui dépeint de futur de l'IA, allant vers l'IA générale (AGI).

App Orchid (2013, USA, $5,5M) propose une solution d'analyse visuelle de données à base de machine learning pour les utilities pour gérer leur grid et sécuriser leurs installations.



Alpiq in Tec (2009, Suisse) propose GridSense, un logiciel d'évaluation des comportements des utilisateurs pour optimiser la gestion de la consommation dans un bâtiment.

Verdigris (2011, USA, $21,6M) détecte la nature des appareils électriques dans un immeuble et fournit le reporting associé, une solution qui rappelle celle de **SmartImpulse** (2011, France).

Stem (USA, $321M) associe plusieurs concepts dans une offre originale : une sorte de grid de production/distribution d'énergie d'origine solaire couplée à un réseau de stockage d'énergie à base de batterie, le tout piloté à base d'IA. L'ensemble regroupe plus de 600 VPP (Virtual Power Plants), des lieux de production d'énergie solaire et une centaine de sites de stockage sur batterie. Le montant de leur financement est étonnant et très élevé.

Engie a de son côté déployé une solution d'optimisation du plan de chauffe des bâtiments en fonction de la météo, de la température extérieure et intérieure, et de la puissance de chauffage disponible. Le tout avec du deep learning associant plusieurs types de réseaux de neurones. Les bâtiments sont regroupés en clusters par type de comportements.

La solution détecte les bâtiments qui dérivent dans leur classe en termes de consommation d'énergie. Cela peut conduire à des recommandations d'optimisation énergétique pour les syndics des bâtiments.

Gestion des déchets

Waste Robotics (2016, Canada) robotisation du recyclage. Premier robot lancé en 2016 pour le tri des matières résiduelles. A base de caméras et de bras mécanique, séparation puis acheminement dans les bonnes zones de valorisation comme le compostage ou la biométhanisation. L'IA utilisée relève ici principalement de la reconnaissance d'images.

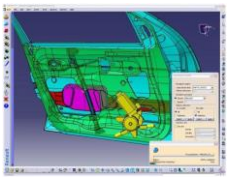
ZenRobotics (2007, Finlande, 14,4M€) est un concurrent de Waste Robotics avec son ZenRobotics Recycler (ZRR). Le bras unique du robot peut sélectionner quatre déchets de types différents en ne se trompant que dans 2% des cas, à raison de plus d'une par seconde. Cela peut notamment trier des déchets de construction. Le tout fonctionne avec des caméras, des capteurs infrarouges, des capteurs de luminosité, des scanner laser 3D (comme des LiDAR), des capteurs haptiques et des détecteurs de métaux

(vidéo).

Industrie

L'industrie manufacturière est probablement le secteur d'activité qui exploite l'IA depuis le plus longtemps, ne serait-ce que dans les systèmes de conception assistée par ordinateurs, dans la simulation et dans la robotique de production, elle-même grosse consommatrice de vision artificielle.

Les systèmes experts sont aussi couramment utilisés dans l'industrie depuis trois décennies, en particulier dans les tâches de maintenance et de réparation, même si cela fait moins l'actualité tandis que le deep learning la monopolise.



conception et simulation
CAO, PLM, VR
mise en situation
simulation d'usine

fabrication
gestion stocks et entrepots
robots de fabrication et assemblage
contrôle qualité par imagerie

exploitation
objets connectés
metering
maintenance prédictive
gestion d'assets

Nous avons déjà rapidement évoqué les robots de production dans la rubrique sur les robots.

Cette partie de cet ebook n'est cependant pas si garnie que cela avec seulement deux pages d'exemples d'application de l'IA. Il faut dire que certaines d'entre elles sont aussi intégrées dans d'autres parties comme les précédentes sur les transports, le BTP et les utilities.

Citrine Informatics (2013, USA, $15,6M) propose une plateforme d'IA généraliste pour les industries de la chimie et des matériaux. Elle semble permettre la collecte et l'exploitation de données structurées et non structurée dans la chimie. C'est visiblement surtout un outil de R&D qui exploite le machine learning avec la visualisation de données et de la segmentation automatique de propriétés de matériaux observées expérimentalement. Elle a notamment permis de conçevoir des poudres d'aluminium utilisables pour de l'impression 3D métallique[503].

Conception

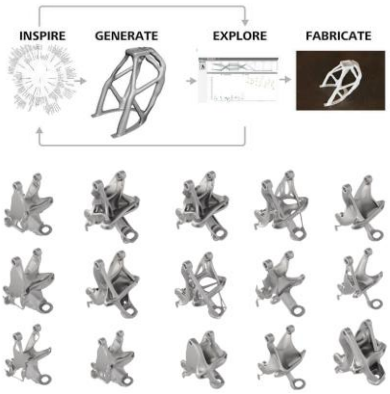
Autodesk a développé en 2017 et commercialisé en 2018 dans Autodesk Generative Design qui est intégré dans sa suite de CAO Fusion 360 une fonctionnalité de conception générative à base de deep learning génératif, ou GAN (Generative Adversarial Networks). Elle permet de générer différentes versions de designs d'objets respectant diverses contraines. Elle exploite une base de formes et d'objets qui est automatiquement adaptée à l'objet en phase de conception.

AUTODESK.

deep learning génératif pour proposer diverses formes dans la conception d'objets respectant un cahier des charge

fonctionnalité Autodesk Generative Design intégré dans l'offre commerciale Fusion 360 Ultimate

anciennement dénommé projet "Dreamcatcher"



[503] Comme la bioinformatique, l'informatique des matériaux est un champ à lui tout seul. Voir notamment The data analytics platform for the physical world de Chris Borg (Citrine Informatics), 2008 (32 slides), Applications of machine learning and artificial intelligence to designing chemicals and materials with the desired properties de Alexander Tropsha, 2018 (64 slides), Machine Learning and Materials Informatics: Recent Applications and Prospects, 2017 (27 pages), Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence, Harvard, 2018 (109 pages) et Machine Learning and Global Optimization for Materials Discovery de Logan Ward, 2017 (52 slides).

Fabrication et contrôle qualité

Dans toutes les usines, il est courant de faire du contrôle qualité des pièces usinées avec des caméras et des solutions de computer vision, comme chez **Foxconn** qui analyse ainsi la qualité de ses cartes électroniques en sortie de bains de soudure. C'est une méthode très courante depuis des années.

un grand classique, le contrôle qualité de production par vision artificielle

source : Nvidia

Arcure (2009, France) développe des solutions de vision artificielle matérielles et logicielles qui détectent les personnes pour des applications de sécurisation d'usines et de chantiers. Elle sécurise notamment les déplacements de robots autonomes.

Scortex (2015, France, $1,8M) conçoit une solution à base de FPGA pour des applications d'inspection et de contrôle qualité à base de traitement de l'image.

Deepomatic (2014, France, $2,3M) utilise ses solutions de vision artificielle pour faire du contrôle qualité visuel dans la production, notamment chez Valeo et Airbus. La startup semble opérer surtout en mode projet et cibler plusieurs marchés verticaux.

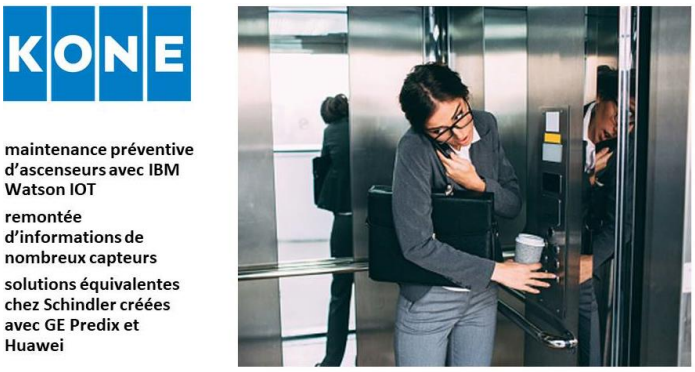
Anodot (2014, Israël, $27,5M) propose une solution de *real-time analytics* qui s'appuie sur du machine learning pour détecter les incidents dans l'industrie en s'appuyant sur des modèles par contraintes sur des modèles de données divers.

Les robots de **Rethink Robotics** (2008, USA, $149,5M) sont aussi intéressants. Ils sont positionnés pour le travail en usine et pour collaborer avec des ouvriers humains. C'est en fait un bras articulé avec un grand nombre de degrés de liberté (vidéo). Manque de bol, la startup qui employait 90 personnes a fermé boutique en octobre 2018, non sans avoir consommé un cash incroyable et déployé ses robots chez quelques industriels aux USA[504]. L'Allemand **Hahn Group** a ensuite acquis les marques et brevets de Rethink Robotics, toujours en octobre 2018.

Dans un positionnement voisin, **Kinema Systems** (USA) développe des solutions de vision 3D pour les robots dans la logistique et la fabrication. Ils avaient gagné le challenge Nvidia de 2018.

Maintenance

C'est le croisement des objets connectés et de l'IA qui génère le plus de nouvelles opportunités de solutions, notamment dans la maintenance préventive et l'optimisation des ressources. La maintenance préventive des ascenseurs fait ainsi appel au machine learning chez les grands ascensoristes du marché tels que **Kone**, **Otis** et **Schindler**. Des capteurs, des bases d'entraînement du deep learning et hop !



maintenance préventive d'ascenseurs avec IBM Watson IOT

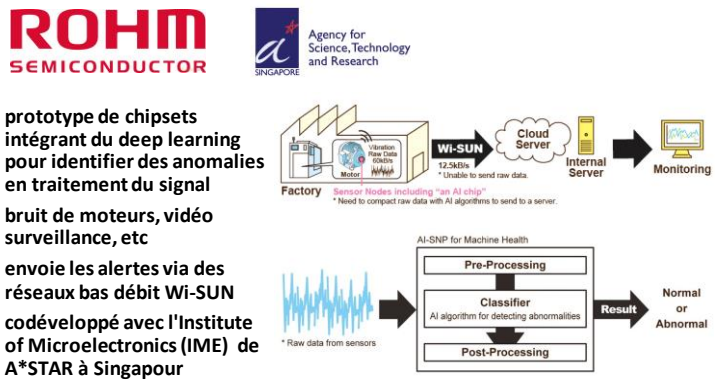
remontée d'informations de nombreux capteurs

solutions équivalentes chez Schindler créées avec GE Predix et Huawei

[504] Ce n'est pas la seule startup de robots à avoir récemment fermé boutique, il y a aussi Mayfield Robotics avec son robot social Kuri et Jibo, un autre robot du même genre, vu au CES 2017. Voir Why Are Robotics Companies Dying? de Ron schmelzer dans Forbes, octobre 2018.

Il existe même des startups spécialisées dans le domaine comme **Uptime** (2016, France, 3M€) qui installe un boitier dans les ascenceurs compatible avec tous les modèles du marché pour récupérer ses données de fonctionnement. Ils ont déjà quelques dizaines de clients.

Le fabricant japonais de composants **Rohm** développe avec codéveloppé avec l'Institute of Microelectronics (IME) de A*STAR à Singapour des composants d'analyse du bruit de moteurs et autres signaux en exploitant du deep learning. Les alertes sont remontées en central via des réseaux de télécommunication bas débit de type Wi-SUN, un concurrent asiatique de LoRA et Sigfox, supporté cependant par STMicroelectronics.



Plus généralement, **Canvass Analytics** (2016, Canada, $5M) fournit une plateforme logicielle complète pour gérer la maintenance prédictive d'une usine. Ils le font notamment dans l'industrie automobile et dans l'agroalimentaire.

Admiral Technologies (2018, France) propose aussi une solution de maintenance prédictive, entre autres, pour les utilities et autour de la captation de données d'objets connectés divers, selon les clients. **Tellmeplus** (2011, France, $4,7M) a l'air positionné sur un créneau voisin.

Objets connectés

Les objets connectés génèrent des données qui sont ensuite interprétées et exploitées par des logiciels à base d'IA. En voici quelques exemples.

Craft.ai (2015, France, $2M) est une jeune startup spécialisée dans l'Internet des objets. Elle permet de créer des solutions logicielles d'orchestration d'objets connectées qui apprennent toutes seules des comportements des utilisateurs et des données environnementales captées par les objets connectés. La solution est commercialisée sous la forme d'APIs destinées aux développeurs d'applications. L'approche est intéressante dans son principe. Reste à trouver un modèle économique solide.

Ubiant (2011, France), basé à Lyon propose une solution matérielle et logicielle de gestion de la maison intelligente, de l'éclairage et de l'énergie qui s'appuie sur du machine learning et sur le Luminion, un objet connecté interagissant avec l'utilisateur via des LED de couleur indiquant si la consommation du foyer est supérieur à celle du voisinage. C'est une offre b2c.

Vivoka (2015, France) a développé Lola, un logiciel de contrôle des équipements de la maison connectée. Elle s'appuie sur une box reliée à Internet qui se pilote via une application mobile et par commande vocale. Le projet lancé sur Kickstarter n'a pas porté ses fruits.

Iqspot (France, 300K€) est une startup bordelaise qui analyse la consommation énergétique des bâtiments et sensibilise ses occupants pour la diminuer. Le tout avec du machine learning. C'est une participation d'IT-Translation.

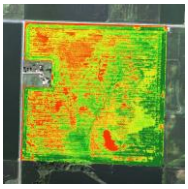
Ezako (2011, France) édite la solution logicielle Upalgo, qui permet d'accélérer l'analyse de données issue notamment des objets connectés couplé à Nector qui s'installe sur des kits Raspberry Pi. Ils s'appuient sur une technologie brevetée de machine learning issue du CNES qui sert à identifier les événements les plus importants dans les données pour identifier les les comportements anormaux. Ils sont déployés chez Safran, Groupe Orange.

Agro-alimentaire

L'agriculture est un autre vaste domaine où l'IA a de nombreuses applications, en particulier en robotique, mais aussi en amont, avec les outils de la télédétection qui s'appuient de plus en plus sur la reconnaissance d'image à base de deep learning et sur l'agriculture de précision qui associe les objets connectés à l'IA ainsi que des robots.

Nous allons explorer ici quelques-unes de ces nombreuses innovations.

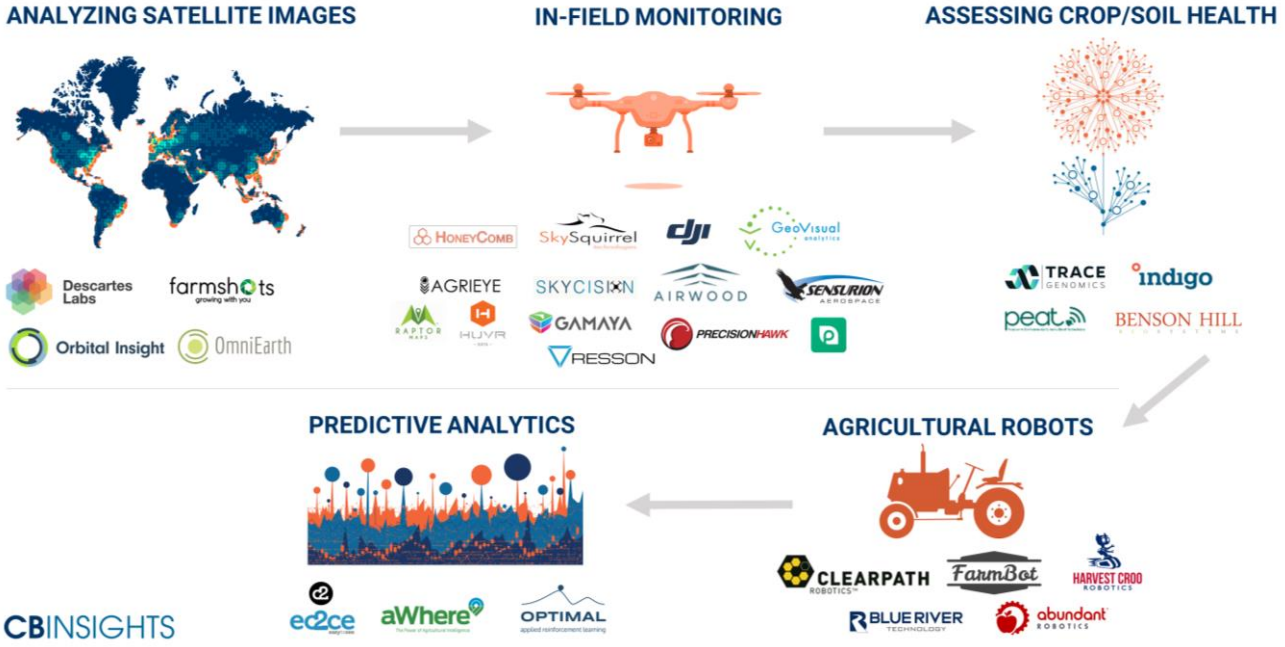
L'IA de l'agriculture utilisée tourne essentiellement autour du traitement de l'image, des données issues d'objets connectés et de la robotique.



télédétection
analyse imagerie drones et satellites
prévisions récoltes

agriculture de précision
exploitation capteurs
planification
optimisation des ressources

robotisation
binage, semis, traitements, récoltes, packaging



CBInsights[505] décrit encore mieux que moi ce paysage avec ce beau schéma qui enchaîne la télédétection de loin (satellite) ou de près (drones), l'évaluation de l'état des champs et récoltes, les robots agricoles et les outils de prévision. Avec un chapelet de startups à la clé, presque toutes des USA. Mais nous n'allons pas toutes les examiner ! En tout cas, il y a fort à parier qu'elles font toutes appel à des briques de l'IA.

Télédétection

L'un des premiers domaines d'application de l'IA dans l'agriculture, que nous avons déjà abordé, touche à la télédétection par satellites et par drones. Il exploite le traitement d'images et notamment les variations dans le temps des observations. Il permet d'évaluer de nombreux paramètres comme la qualité des terrains, leur hydratation et le niveau de prévision des récoltes, à l'échelle de son ex-

[505] Source : AI, Robotics, And The Future Of Precision Agriculture, juillet 2017.

ploitation aussi qu'à l'échelle globale, ce qui permet d'anticiper les cours de vente de sa production voire de les optimiser.

En plus de **Descartes Labs** et **Orbital Insight**, déjà cités dans la partie sur la télédétection générique, nous pouvons citer également…

Farmshots (2014, USA, acquis par le Suisse Syngenta en 2018) qui détecte les maladies, et les besoins en nutriments des plantes par l'analyse d'images satellites et de drones. C'est lié à l'agriculture de précision qui vise à, par exemple, focaliser les pesticides ou l'arrosage là où ils sont le plus nécessaires. En règle générale, ces analyses utilisent un spectre électromagnétique large, comprenant notamment l'infrarouge qui détecte le mieux les variations d'états dans les plantes. **Flurosat** (2016, Australie, $2,4M) est sur le même créneau.

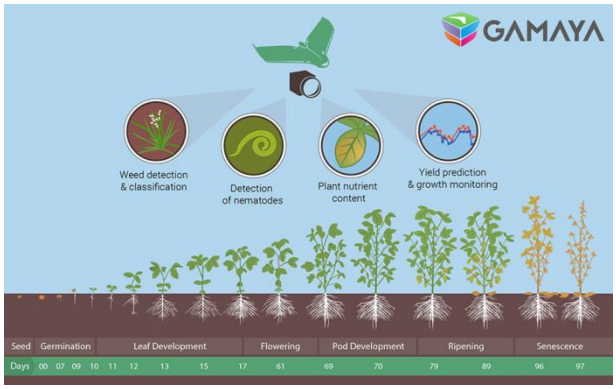
Farmlogs (2012, USA) exploite l'imagerie aérienne pour suivre l'état des champs de céréales et de coton et les traiter convenablement.

OmniEarth (2014, USA, $5,2M, acquis par EagleView Technologies en 2017) est focalisé de son côté sur l'analyse hydrique des terrains.

HoneyComb (2012, USA, $415K) analyse les images provenant de drones d'observation. Ils proposent le drone qui va avec, le AgDrone System.

Agrieye (2016, Ukraine, $150K) fournit comme HoneyComb une solution intégrant un drone d'observation et un service en cloud d'analyse des terrains.

Skycision (2015, USA, $1,5M) s'appuie de son côté sur les drones du Chinois DJI et les complète par ses services en cloud.



Gamaya (2015, Suisse, $4M) couvre aussi bien les grandes surfaces agricoles que les petites exploitations. Comme pour tous les acteurs de ce marché, il s'agit d'analyser les terrains pour optimiser et cibler l'usage de pesticides et d'engrais.

AirWood (2014, Inde) utilise des drones équipées de caméras multispectrales. Le reste est classique. Même chose chez **GeoVisual Analytics** (2000, USA, $1,2M).

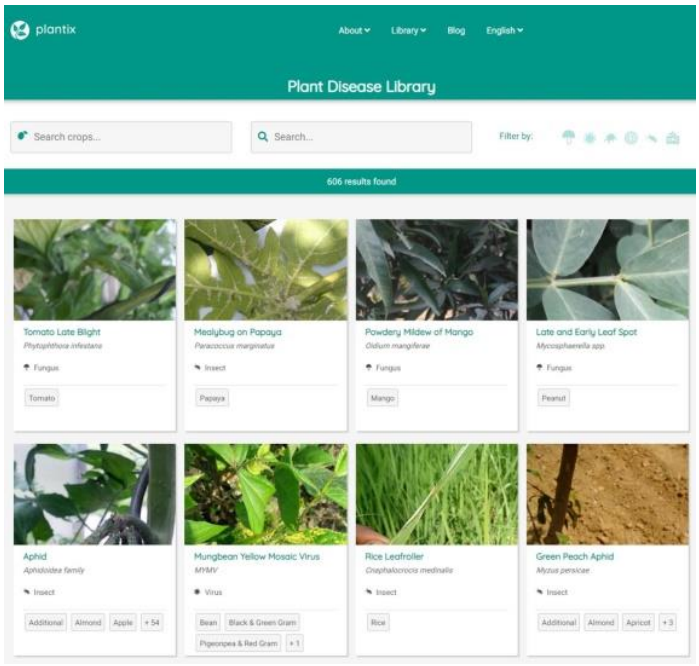
Agriculture de précision

L'agriculture de précision exploite les données issues de l'imagerie aérienne et de capteurs pour assurer au mieux le rendement des récoltes, l'emploi de pesticides et d'engais et l'arrosage. C'est donc une affaire de consolidation de données et de machine learning permettant de détecter des anomalies et de faire des recommandations d'actions et des prévisions de récoltes.

Bowery Farming (2015, USA, $31M) propose son système d'exploitation BoweryOS qui s'appuie sur de la vision artificielle et du machine learning pour suivre l'état de plants et optimiser leur croissance en diminuant le besoin en eau.

PEAT (2015, Allemagne) a développé l'application mobile Plantix qui exploite de simples photos de végétaux prises par des smartphones pour identifier les maladies ou parasites qui les affectents.

L'application fournit des recommandations diverses pour le traitement des plantes. L'application est gratuite et la startup espère exploiter les données récupérées, un modèle économique toujours difficile à mettre en place.



Benson Hill BioSystems (2012, USA, $94,7M) a développé la plateforme CropOS servant à prédire le rendement de récoltes en fonction de différentes caractéristiques des plantes, comme leur capacité à optimiser la photosynthèse via leur ADN. La société a aussi créé un outil d'édition de gênes CRISP 2.0 censé être plus efficace que le très connu CRISP-Cas9.

Prospera Technologies (2014, Israël, $22M) utilise de l'imagerie des champs, des capteurs et des sondes météorologiques pour déterminer leur santé, le niveau de nutriments et d'hydratation.

Bilberry (2016, France) a créé une solution de scan de la végétation présente au sol en temps réel, attachée aux engins de pulvérisation, pour ne déposer du désherbant que là où cela est nécessaire.

Copeeks (2016, France) commercialise des capteurs divers qui permettent de suivre l'état des récoltes. Dont un capteur vidéo qui analyse l'état des champs. Il exploite le Microsoft Custom Vision SDK (vidéo).

J'avais sinon croisé un très grand nombre de drones agricoles chinois au CES 2018. Ceux-ci étaient surtout destinés à l'épandage sélectif de pesticides.

Robots agricoles

L'IA intervient surtout dans les techniques de robotisation d'exploitation. Comme partout ailleurs, les robots de l'agriculture sont très spécialisés. Certains s'occupent des animaux comme pour la traite des vaches mais l'essentiel est lié au cycle de vie des récoltes allant du semis aux récoltes.

Les robots présentent l'avantage théorique de permettre des économies de main d'œuvre sur des travaux qui sont en général pénibles et saisonniers.

Reste à faire en sorte que cela soit rentable, les robots transférant des dépenses d'exploitation (salaires) vers des dépenses d'immobilisation (investissement dans les robots) sauf s'ils sont loués.

Nombre de robots agricoles sont surtout des projets de laboratoires de recherche qui n'ont pas pour autant abouti des années plus tard à des produits industriels. C'est le cas du projet **CASC** (Comprehensive Automation for Specialty Crops) de l'Institut de robotique de l'Université Carnegie Mellon, focalisé notamment sur les récoltes de pommes et d'autres arbres fruitiers et dont les vidéos datent de 2012. Ils planchaient même sur des robots d'estimation de taille de récolte de fruits.

Le projet a été financé par le département de l'Agriculture fédéral US (USDA) à hauteur de $10M entre 2008 et 2012. Y participait l'industriel John Deere qui ne semble pas avoir transformé cela en

robots industriels. Il est en effet encore difficile de créer des robots fiables et à des coûts raisonnables pour ces tâches.

La manipulation directe des fruits et légumes est une sacrée paire de manche. Cela fait une bonne vingtaine d'années que des robots sont mis au point pour les récoltes de fruits et légumes en tout genre : ramassage de fruits dans les arbres comme les pommes, de melons, de tomates, de fraises, de concombres[506], d'asperges et même la récolte et la découpe de fleurs.

Ces tâches sont complexes à mener. Les robots doivent détecter les fruits et légumes de taille et formes diverses, qui sont souvent cachés derrière des feuilles ou des branches. Ils doivent ensuite les récupérer avec précaution, sans les abimer puis les placer dans un récipient mobile. En extérieur, les robots doivent si possible résister aux intempéries, une contrainte que l'on n'impose pas aux robots dans les usines.

Une méthode consiste à changer la forme des arbres pour les adapter à la récolte des pommes ! Un peu comme un utilisateur d'ordinateur ou de mobile s'est habitué aux idiosyncraties fréquentes de ces appareils.

La PME française **Carré** a conçu Anatis 2 en 2014 (*ci-dessous à gauche*), un robot bineur dédié aux cultures maraichères et équipé de caméras en tout genre pour se mouvoir et analyser le terrain (vidéo). Basées à La Roche sur Yon, la PME est malheureusement en procédure de sauvegarde d'entreprise depuis septembre 2016. Elle faisait près de 11M€ de CA en 2016, mais avec un déficit visiblement chronique de plus d'un million d'Euros. L'entreprise qui compte 87 salariés produisait historiquement de l'outillage agricole traditionnel. Il est difficile d'innover dans une PME traditionnelle !

Le français **Naïo Technologies** (2011, France, 3,5M€) a créé un autre robot de binage, le Oz, qui a été produit en quelques dizaines d'exemplaires (vidéo et *ci-dessus à droite*). Il permettrait de diminuer l'usage des produits phytosanitaires. Le français **SITIA** a créé un tracteur autonome (vidéo). **EcoRobotix** (2011, Suisse, $3M) a créé un robot de désherbage (vidéo) qui a un peu plus de chances d'être commercialisé que le robot **Weedmaster**, un concept du designer industriel suisse Fabian Zimmerli qui n'a pas été industrialisé.

Le robot de surveillance et de désherbage Thorvald 2 de **Saga Robotics** (2016, Norvège) permet d'éviter de faire appel à des pesticides (vidéo, *ci-dessous à droite*). Il semble qu'il s'agisse d'un projet de recherche. Il semble qu'il en soit aussi de même d'un autre robot de désherbage, le AgBot 2, de l'institut de recherche australien **QUT Research** (vidéo, *ci-dessous à gauche*).

[506] Vous rigolez, mais un japonais, Makoto Koike, a créé un système de tri de concombres animé par une application développée avec TensorFlow (vidéo et source de l'information) ! Il utilise de la vision artificielle et détecte plusieurs paramètres : la couleur, la taille, la forme, des défauts des concombres analysés.

Vitirover (2007, France) a développé un robot de désherbage des vignobles et autres types de terrain qui mesure aussi la température et l'hygrométrie au sol et détecte diverses pathologies. Il est en plus alimenté par énergie solaire.

BlueRiver Technology (2011, USA, $30,3M, acquis par John Deere en 2017) conçoit un robot d'épandage sélectif de pesticide (vidéo). Mais là encore, il n'a pas l'air d'être commercialisé.

CNH Industrial a créé un concept de tracteur autonome pour la récolte de céréales (vidéo) qui travaille la nuit sans broncher. Il est pilotage par tablette pour la programmation des tâches et des terrains à traiter.

Nombre de robots de récole de fruits sont ainsi encore généralement du domaine de l'expérimentation, comme pour les concombres (vidéo), les pommes destinées à la production de cidre chez SFM Technology (vidéo) ou les poivrons (vidéo).

Une startup de San Diego, **Vision Robotics** (1999, USA), s'attaque à la récolte de raisins, histoire de remplacer les travailleurs immigrés du Mexique qu'il est plus difficile de faire traverser la frontière (vidéo qui date de 2012) !

La spin-off du laboratoire de recherche SRI de Menlo Park **Abundant Robotics** (1016, USA, $12M) a créé un robot expérimental de récolte de pommes dans l'Etat de Washington, l'un des plus gros producteurs de pommes aux USA (vidéo), avec la Pennsylvanie (« Johnny Appleseed »). Il les aspire avec une sorte de ventouse (*ci-dessous à droite*)[507].

La culture et la récolte des champignons fait l'objet de recherches à l'**Université de Warwick** au Royaume Uni. Leur robot expérimental de récolte sait détecter les champignons arrivés à maturité pour les récupérer, via l'analyse de leur taille par reconnaissance d'images. Ces robots doivent travailler 24h sur 24 et dans des environnements sous-terrains pas très hospitaliers.

[507] Cette intéressante présentation décrit différentes méthodes de récolte de pommes : Agricultural Robotics: Opportunities, Challenges and Perspectives de Qin Zhang, 2015 (23 slides).

Au Japon, les chercheurs de l'**Université d'Okayama** planchent depuis plus de 25 ans sur les robots de récolte de tomates, concombres, raisins et fraises. Une analyse spectracle dans le proche infrarouge des fruits est réalisée pour détecter ceux qui sont mûrs (sorte de Scio à usage professionnel). **Panasonic** a aussi développé son propre robot de récolte de tomates fin 2017 mais il a encore tout l'air d'être un simple prototype (vidéo). Il doit récolter 10 tomates à la minute, soit le rythme d'un ouvrier agricole mais pas mieux.

Four Growers (2017, USA, $120K) conçoit aussi des robots de récolte de tomates et de raisins dans les serres mais la jeune startup est encore loin de la commercialisation. **Harvest Croo** (2013, USA, $3M) met au point un robot de récolte de fraises, qui n'est donc lui aussi visiblement pas encore commercialisé (vidéo).

On peut ajouter à cet inventaire à la Prévert le **Prospero**, un robot de semis à cinq pattes qui est toujours un prototype (*ci-dessus à gauche*).

Le robot ramasseur de fraises SW 6010 de l'Espagnol **Agrobot** (vidéo 1, vidéo 2) a l'air opérationnel, avec cueillette robotisée mais intervention humaine pour remplir les barquettes dans sa première version et entièrement automatique dans la seconde (*ci-dessous à gauche*). **Traptic** (2016, USA, $3,4M) est sur le même créneau avec un robot qui n'a pas l'air d'être encore commercialisé.

L'américain **Blue River Technologies** (2011, USA, $30,3M) propose de son côté un système robotisé de culture de laitues contenant une bardée de capteurs, dont certains sont 3D, pour optimiser l'entretien de laitues ou de plants de mais (*ci-dessous à gauche*). Il a aussi lancé des tracteurs de semis de coton dont seul l'outillage, mais pas le pilotage, est robotisé, avec des caméras qui détectent les mauvaises herbes et un système qui dépose dessus de l'herbicide (*ci-dessous à droite*).

Les tracteurs autonomes sont plus faciles à mettre au point car ils reprennent des techniques relativement éprouvées de véhicules autonomes. C'est le cas du Bonirob des allemands **Bosch** et **Amazonen-Werke** qui est un tracteur autonome modulable (*ci-dessous*) et peut servir notamment au binage de la terre. Mais il n'a pas l'air d'être encore commercialisé à grande échelle pour autant.

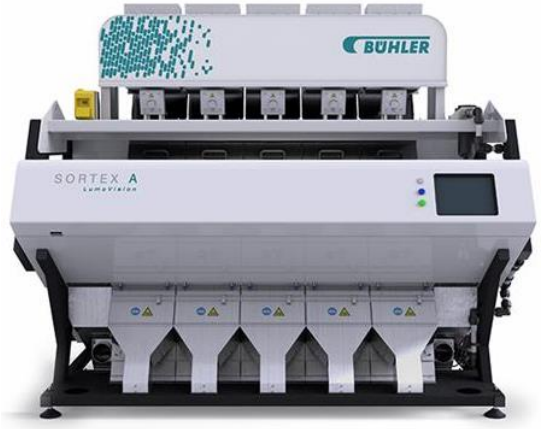
L'Américain **ATC** (Autonomous Tractor Company) se positionne comme le Tesla des tracteurs en les électrifiant entièrement. Ils ont aussi conçu un tracteur autonome électrique… à l'état de concept (*ci-dessous à gauche*).

L'Américain **Harvest Automation** (2009, USA, $33,5M) a créé un petit robot, le HV-100 qui déplace des pots, une tâche pas trop complexe voisine de celle des robots d'entrepots. On en revient aux choses simples !

A terme, on verra apparaître des fermes où l'ensemble des processus sont robotisés, surtout pour les cultures sous serres. Dans la nature, les robots doivent composer avec des terrains par toujours réguliers.

Après ces quelques recherches de robots agricoles, je me rends compte finalement que l'on en est à peu près au même stade que pour les robots humanoïdes : les démonstrations et effets d'annonce sont nombreux, mais les réalisations concrètes opérationnelles le sont bien moins. Cela ne veut évidemment pas dire que cela ne marhera jamais mais que la mise au point de ces robots agricoles dans des conditions économiques satisfaisantes est encore un long chemin semé d'embuches.

En amont des récoltes, l'IA peut jouer un rôle dans le contrôle qualité, en particulier via de l'inspection visuelle. C'est ce que réalise la société Suisse **Buhler** qui intègre des capteurs d'images dans ses systèmes d'inspection de céréales. Ils exploitent notamment la partie proche UV et proche infrarouge du spectre électromagnétique dans leur machine Lumovision et détectent les défauts des grains pouvant révéler des pathologies diverses comme l'aflatoxine, un champignon toxique qui peut se développe sur les grains récoltés. Le client peut éliminer des lots complets ou sélectionner automatiquement les grains à éliminer.

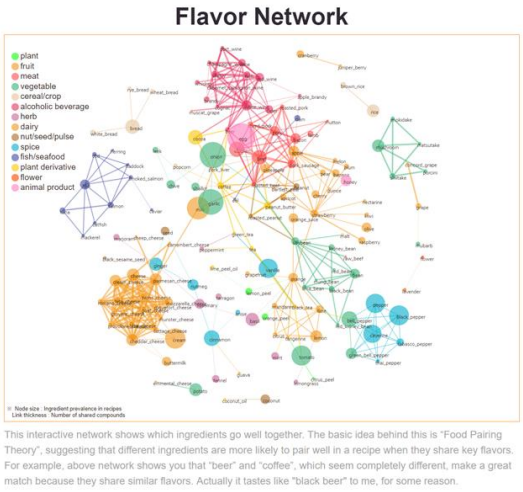


Alimentaire

L'IA peut être utilisée dans plein de registres dans les industries agro-alimentaires. Elles réexploitent des techniques vues dans la rubrique sur l'industrie concernant la fabrication de produits en usine, ne serait-ce que pour la robotisation et le contrôle qualité. C'est ce que propose **Dataswati** (2016, France), une startup qui utilise une IA multimodale exploitant de l'imagerie et des données de capteurs divers (humidité, …) pour faire du contrôle qualité en particulier dans les chaînes de production agro-alimentaires. Ils démontraient aux Universités d'Eté du MEDEF 2018 leur système pour évaluer la qualité de différentes tranches de saucisson ! Voilà du concret !

Quelques autres applications spécifiques plus ou moins exotiques voient le jour.

Un certain Yoshiki Ishikawa a réalisé une cartographie du goût et des compatibilités entre ingrédients réalisée[508] (*ci-contre*). **Foodpairing** (Belgique) tire parti de l'analyse de 2000 ingrédients courants et de ses 8000 molécules d'aromes pour également savoir lesquelles pouvaient être associées et satisfaire nos palais délicats. Les clients sont les industries agro-alimentaires et les chefs cuisiniers. Ce sont peut-être des algorithmes de ce genre qui sont utilisés par le brasseur **Carlsberg** qui ambitionne d'utiliser l'IA pour découvrir de nouvelles saveurs de bière[509]. En attendant, ils pourraient recréer la Hardy's Ale écossaise qui a disparu du circuit il y a une dizaine d'années !



Yummly (2009, USA, acquis par Whirlpool en 2017) est une variante de Marmiton.org qui vous fait des propositions culinaires en fonction de vos goûts, du lieu, du moment dans la journée et de la saison. L'application a crunché deux millions de recettes qui ont été crunchées par une IA non précisée[510].

Cette rubrique ne serait pas complète sans faire un tour du côté des robots cuisiniers. Les robots ménagers grand public qui rapent les carottes n'ont rien de robots. Il faut chercher plus loin pour trouver des machines qui ressemblent de plus près à des robots. On en trouve qui font le café, mais ce sont surtout des automates.

Commençons avec **Ekim** (2012, France, 2,2M€) avec son robot qui cuisine des pizzas, le Pazzi. Il comprend deux bras articulés 6 axes (vidéo) est s'intègre dans une cuisine complète pour gérer le processus de préparation et de cuisson de la pizza. C'est compliqué à mettre au point mais cela relève de l'état de l'art, pas de le faire repousser. La pizzeria autonome sera à 500K€, ce qui nécessitera d'avoir un bon flux de clients pour rentabiliser l'investissement.

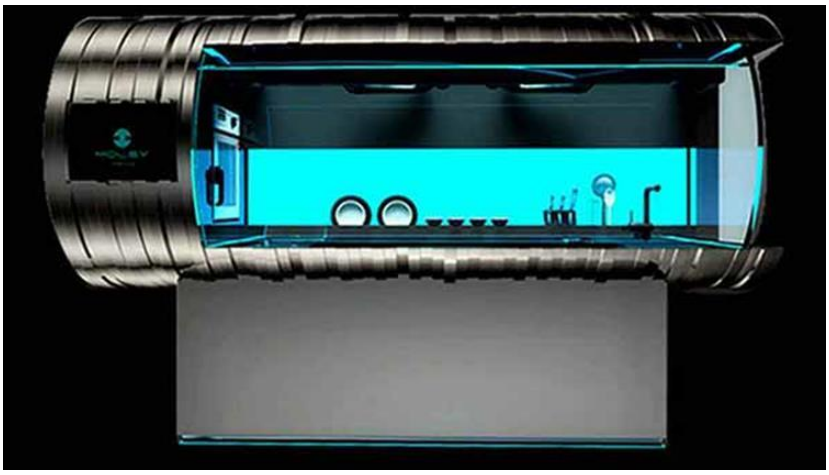
Et puis voici une étonnante startup, **Zume Pizza** (2015, USA, $423M) qui est au départ un service de foodtrucks de pizzas californien. Ses foodtrucks sont alimentés en ingrédients comestibles et en IA pour prédire la demande des consommateurs et automatiser une partie de la production des pizzas alors que le camion est en route (vidéo). Dingue. Mais au vu des vidéos, il semble que les pizzas ne soient pas préparées dans les camions.

[508] Voir Flavor Network et Flavor network and the principles of food pairing, 2011 ainsi que Food-bridging: a new network construction to unveil the principles of cooking, 2017, et Data Science in Cooking de Stephanie Dodson, 2017 (20 slides), qui fait du clustering de styles régionaux de cuisine.

[509] Vu dans Une intelligence artificielle pour aider à la création de nouvelles bières de Pierrick Labbé, juillet 2018.

[510] Ces trois exemples proviennent de L'intelligence artificielle s'invite dans nos assiettes, par Morgane Thual, mars 2017.

Le projet le plus intéressant est celui de **Moley Robotics** (2015, UK, $1,2M), un robot conçu par Mark Oleynik dont le lancement était prévu en 2018. Le prototype a été développé avec Shadow Robots (1987, UK). Il comprend deux bras articulés capables de manipuler les principaux outils de la cuisine. Les gestes ont été appris au robot en captant ceux d'un chef, Tim Anderson (vidéo 1, vidéo 2).

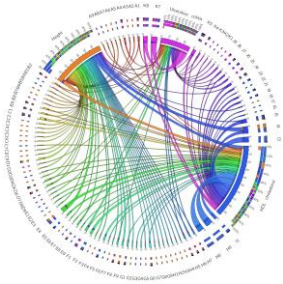


Santé

La santé est l'un des marchés verticaux les plus attirants pour les startups de l'IA avec celui de la finance et du commerce.

L'IA est notamment utilisée dans trois principaux domaines : celui du diagnostic avec diverses medtechs dont l'imagerie médicale ou la génomique et la médecine prédictive, celui des thérapies avec les outils pour les biotechs ou les robots chirurgiens, et enfin, sur la gestion des systèmes de santé et des soins dans la durée.

diagnostics
imagerie médicale
ECG, EEG
médecine prédictive
génomique
télémédecine



thérapies
drug discovery
criblage de molécules
simulation du vivant
radio/chimio ciblée
robots chirurgiens

systèmes de santé
suivi post-opératoire
prévention des erreurs
réduction des risques
prévision de coûts

Ce qui suit n'est probablement qu'une liste très partielle des startups de ce secteur d'activité florissant.

grande vague de startups dans l'IA et la santé !

Diagnostics

L'aide au diagnostic est probablement le domaine de la santé où l'IA a le plus prospéré ces dernières années. Le deep learning et les réseaux de neurones convolutionnels sont notamment omniprésents dans l'interprétation de l'imagerie médicale. Il rend accessibles les connaissances des spécialistes aux médecins généralistes, ce qui est particulièrement utile dans les déserts médicaux et dans les pays émergents. Ils permettent d'augmenter la portée de certains dépistages pour éviter que des pathologies se développent silencieusement. C'est une des formes de la médecine prédictive. Ce sont probablement aussi les applications les plus éprouvées de l'IA dans la santé.

Les solutions d'imagerie utilisent à peu près toutes les mêmes techniques et solutions logicielles mais sont paramétrées de manière différente selon les pathologies recherchées et avec des jeux d'entraînement spécifiques. Elles vont bien au-delà des techniques simples d'augmentation de contraste. Elles détectent des formes particulières, des densités spécifiques et réalisent aussi des mesures précises et quantitatives.

Elles peuvent aussi comparer avec précision des images dans le temps pour quantifier l'évolution d'une pathologie. Elles utilisent généralement différentes formes de réseaux convolutionnels, avec au moins un premier qui détecte des formes dans l'image et un autre qui les labellise une par une après détourage.

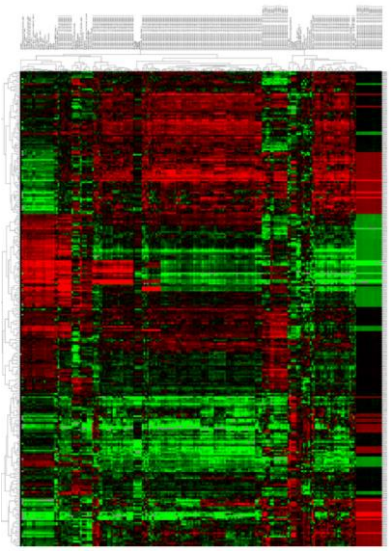
Pour être entraînées, ces solutions doivent évidemment exploiter des bases d'images déjà tagées par des spécialistes et issues de réseaux de laboratoires, cliniques et hopitaux. Ces données ne sont pas forcément ouvertes et les startups doivent monter des partenariats ad-hoc pour les récupérer. Très souvent, les startups françaises se sourcent aux USA pour récupérer ces bases de données[511].

Les rapports d'analyses de ces systèmes sont aussi rédigés avec des textes en clair, ce qui relève du traitement du langage[512] avec des réseaux à mémoire (LSTM) et même des réseaux génératifs (GAN).

[511] A noter l'existence de **TeraRecon** (1999, USA), un fournisseur de bases d'imagerie médicale qui permet d'entraîner des IA. Ils disposent de centaines de millions d'examens y compris de l'imagerie 3D. Leur filiale WIA Corporation issue de l'acquisition de McCoy Medical Technologies en 2017 propose l'accès à des algorithmes d'IA via une plateforme et un jeu d'APIs. C'est en gros un intermédiateur d'algorithmes de traitement d'images.

[512] Une méthode est décrite dans On the Automatic Generation of Medical Imaging Reports 2018 (10 pages).

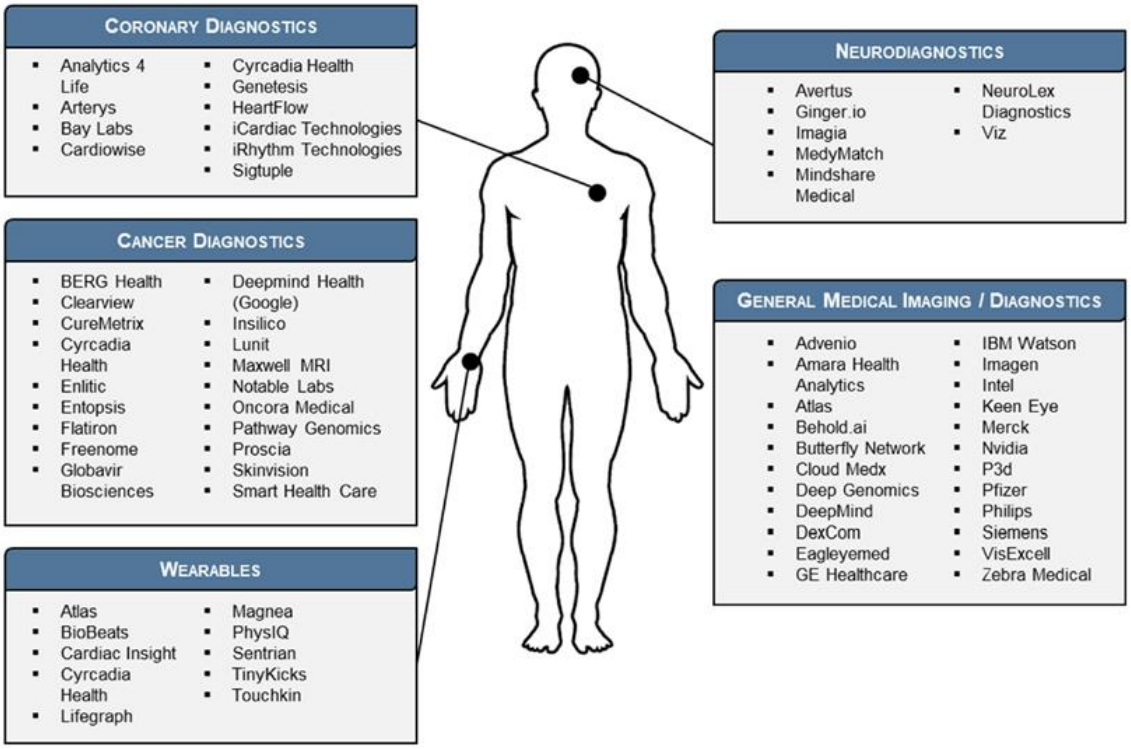
Une nouvelle discipline née en 2012 complète l'analyse d'images classique avec tagging des tumeurs ou pathologies découvertes par des réseaux convolutionnels, les « **radiomics** »[513].



Il s'agit d'une méthode différente d'analyse des images qui génère des données quantitatives sur des « features » de ces images. En gros, c'est de l'analyse de données macro sur les images. Elle permet d'identifier des caractéristiques pathologiques qui ne sont pas détectées par l'analyse visuelle classique d'images et par réseaux convolutionnels.

Elle peut-être aussi croisée avec des données autres comme celles du génome ou des données phénotypiques pour identifier des corrélations. Cela génère des visualisations étonnantes comme *ci-contre*, avec dans les lignes, l'apparition de ces « features » dans les images en en colonne, des patients ayant ici un cancer du poumon.

Quasiment tous les pans de l'imagerie médicale sont touchés par le deep learning pour la détection de pathologies. Je vais les décrire un par un avec quelques solutions du marché. Il faut cependant apporter une nuance : si les prouesses de l'IA sont largement commentées dans ce domaine, les solutions associées sont encore loin d'être déployées. Les professions de santé sont naturellement conservatrices. Les bénéfices de ces solutions n'ont pas encore été évalués à grande échelle. Les régulateurs de santé sont aussi lents à la détente. Donc, il faut bien séparer les effets d'annonces de l'adoption de ces technologies.



[513] Pour en savoir plus sur ce vaste sujet, voir Radiomics the process and the challenges, 2012 (15 pages), Radiomics for outcome modeling: state-of-the-art and challenges de Mathieu Hatt (INSERM), 2018 (60 slides), Radiomics Images are More than Meets the Eye, de Ralph Leijenaar, 2017 (197 pages), Machine Learning Applications for Radiomics Towards Robut Non-Invasive Predictors in Clinical Oncology de Chintan Parmar, 2017 (187 pages) et Radiomics applied to lung cancer: a review, 2018 (12 pages).

Par contre, nombre d'études illustrent le fait que ces systèmes génèrent des résultats plus fiables que ceux des spécialistes[514].

D'autres montrent que l'association IA plus spécialiste fournit des résultats plus fiables que les spécialistes ou l'IA utilisés isolément[515] ce qui illustre le fait que l'IA est un outil qui complète les spécialistes plus qu'il n'a vocation à les remplacer.

Comme l'illustre le schéma *ci-dessous*[516], le nombre d'acteurs de ce marché est significatif et nous n'en citerons ici que quelques-uns.

Ophtalmologie

L'IA est déjà largement utilisable pour l'analyse du fonds de l'œil afin de détecter diverses pathologies comme les microrétinopathies diabétiques ou les débuts de glaucome ou de dégénérescence maculaire (DMLA). En général, les ophtalmologues dilatent la pupille des patients avec un collyre avant l'examen du fond de l'oeil[517]. Ils utilisent ensuite un dispositif optique qui éclaire le fond de l'œil et récupère une image. L'inspection habituelle est réalisée par l'œil de l'ophtalmologue ou par une caméra qui affiche ensuite le résultat sur un écran en temps réel.

Les solutions à base de deep learning qui caractérisent les pathologies occulaires sont, à ce jour, plus performantes que les spécialistes. Mais il ne faut pas s'en étonner : lorsque l'on entraîne une IA avec le savoir d'un grand nombre de spécialistes d'un métier, il est normal qu'elle soit plus performante qu'un spécialiste isolé[518].

La détection peut concerner le **glaucome,** une des principales causes de cécité dans le monde qui est détectée trop tardivement dans la moitié des cas. Une expérience a été menée avec **IBM Watson** en Australie en 2016 avec un entraînement basé sur l'exploitation de 88 000 fonds de rétines[519].

Elle porte aussi sur la **rétinopathie diabétique**, qui correspond à une altération des micro-vaisseaux sanguins qui irriguent la rétine au fond de l'œil[520]. Un premier système de ce genre a obtenu l'agrément de la FDA aux USA en avril 2018, le IDx-DR de la startup **IDx** (2010, USA,$11,5M). Celui-ci avait été testé sur 900 patients avec de très bons résultats en diagnostics négatifs et positifs[521]. C'est un dispositif médical associant un matériel d'exploration optique et le logiciel, bref, une solution de bout en bout[522]. Une performance équivalente a été réalisée par des chercheurs chinois et avec un entraînement sur 80 000 fonds de l'œil et un taux de réussite de 91%[523]. Ce genre d'examen est aussi proposé par **AIvision.healh** (France) ainsi que **AiScreenings** (2017, France). **Eyenuk** (2010, USA, $5,9M) propose la détection de rétinopathie diabétique avec son logiciel EyeArt exploitable à partir d'un simple smartphone et d'un accessoire optique léger.

---

[514] Voir par exemple China Focus: AI beats human doctors in neuroimaging recognition contest, juin 2018.

[515] Voir AI-Human "Hive Mind" Diagnoses Pneumonia, septembre 2018.

[516] Issu de The Next Generation of Medicine: Artificial Intelligence and Machine Learning de TM Capital, novembre 2017 (25 pages).

[517] A terme, on pourra peut-être se passer de cette dillatation grâce à des caméras adaptées. Voir par exemple Un fond d'œil obtenu sans dilatation de l'iris grâce à une caméra bon marché, avril 2017.

[518] Voir A.I. Equal to Experts in Detecting Eye Diseases, août 2018.

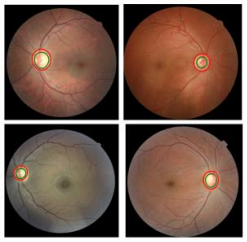
[519] Voir Watson's detective work could help stop the silent thief of sight, février 2017.

[520] Voir How AI Enhances & Accelerates Diabetic Retinopathy Detection, un livre blanc de **Cognizant,** février 2018 (16 pages).

[521] La solution et l'état de l'art sont est très bien documentés dans AI is poised to revolutionize medicine. An overview of the field, with selected applications in ophthalmology de l'American Academy of Ophtalmology, 2018.

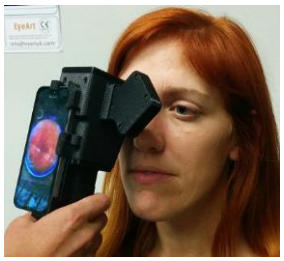
[522] Mais les médias ont toujours tendance à généraliser. Comme ce titre : Aux États-Unis, une intelligence artificielle peut désormais faire des diagnostics médicaux, dans Top Santé, en avril 2018. Cela sous-entend que le système est généraliste alors qu'il est hyperspécialisé sur une seule pathologie.

[523] Voir Chinese researchers develop AI technology for screening diabetic retinopathy, juin 2018.



glaucome

rétinopathie diabétique



rétinopathie diabétique

La filiale UK d'Alphabet **DeepMind** est capable de détecter une cinquantaine de pathologies oculaires d'un seul coup avec 94% de précision, en association le Moorfields Eye Hospital de Londres, dont la **dégénérescence maculaire** (DMLA) ainsi que le **décolement de la rétine**. L'outil s'appuie sur la réalisation de scans rétiniens en 3D. Il classifie les pathologies par gravité en indiquant le degré d'urgence du traitement[524]. Le système utilise une combinaison de plusieurs réseaux de neurones convolutionnels entraînés avec 14 884 scans issus de 7621 patients. Les déploiements sont prévus d'ici 2023 !

ORL

La détection des **pathologies de l'oreille** comme les oreillons et les éclatements de tympans par analyse du tympan est un autre domaine où l'IA a toute sa place pour accélérer les diagnostics[525].

i-nside (France), fondée par le docteur Laurent Schmoll, utilise un otoscope conçu par ses soins et fabriqué par l'Allemand Stortz[526] pour smartphone et une solution de deep learning entraînée avec 250 000 images, via le service de **Clarifai**[527] qui peut fonctionner en mode embarqué depuis 2017 et ainsi être accessible aux praticiens dans le monde qui n'ont pas accès à Internet sur leur mobile (vidéo).

Cardiologie

La cardiologie fait appel à plusieurs outils de diagnostic : les échographies, les radios, les IRM, et les ECG (électro-cardiogrammes). Chacun joue son rôle dans le diagnostic ou la prévoyance. Chacun des résultats de ces analyses peut être exploité par l'IA et du deep learning.

---

[524] Voir L'intelligence artificielle DeepMind, de Google, peut détecter 50 maladies oculaires aussi bien que votre ophtalmo, de Hugo Jalinière, août 2018 et L'IA DeepMind détecte les problèmes oculaires aussi bien qu'un expert humain, août 2018.

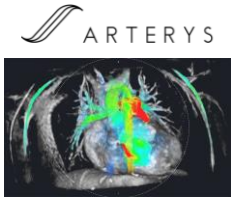
[525] Voir Cutting Edge Technologies in Otology Field, 2017 (5 pages).

[526] Le fondateur de i-nside, a déposé en 2013 le brevet de son Smart Scope, un petit objectif qui relie les endoscopes aux smartphones. Le brevet est utilisé sous licence par l'Allemand **K. Storz** depuis 2015.

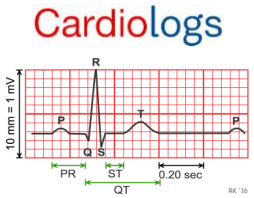
[527] La solution d'I-nside est une étude de cas documentée par Clarifai sur https://clarifai.com/customers/i-nside.

L'analyse d'échographies à base d'IA se trouve notamment chez **Bay Labs** qui fournit aussi ses propres échographes portatifs (2013, USA, $5,5M) et **Arterys** (2011, France, $43,72M) [528] ainsi que chez **DIA Imaging Analysis** (2009, Israël, $10,1M).

échographies cardiaques



athéromes



ECG

Behold.ai (2015, USA, $255K) a développé une solution d'analyse d'imagerie médicale pour aider les radiologues à faire leur diagnostic. Le système compare les images de radiologie avec et sans pathologies pour détecter les zones à problèmes, comme les nodules et autres formes de lésions.

Analytics 4 Life (2012, Canada, $29M) développe son système d'imagerie cardiaque original CorVista pour la détection de pathologies coronariennes qui s'appuie sur un ECG à six électrodes fonctionnant à 8000 Hz pendant 3,5 minutes et réalisée au repos.

Il utilise une technique de « Phase Space Tomography » qui permet de reconstituer une image en 3D du cœur[529] à partir des signaux captés et de leur déphasage, pour faire une sorte de triangulation. Le tout s'appuie sur de l'IA, mais la technique n'est pas du tout précisée dans leur littérature. Le traitement est réalisé dans le cloud mais le résultat est semble-t-il couplé à ceux d'une tomographie aux rayons X. Les études cliniques sont en cours.

HealthMyne (2013, USA, $26,4M) propose aussi un logiciel généraliste d'analyse de radios qui produit des rapports quantitatifs sur certaines observations.

Cardiologs (2014, France, $7,9M) est une startup française qui développe une solution logicielle en cloud d'analyse les données des ECG réalisées selon les règles de l'art avec plusieurs patches d'électrodes (4 sur les membres, 6 sur le thorax) en cabinet médical, par des infirmiers ou des docteurs. Les résultats sont fournis sur une interface web. Elle s'appuie sur des méthodes de machine learning exploitant des réseaux de neurones convolutionnels avec apprentissage supervisé (ConvNets). Côté cloud, ils utilisent comme nombre de startups de l'IA les ressources de Google Tensorflow. Cela permet de fournir une réponse en quasi-temps réel. Ils ont entraîné leur système avec des bases d'ECG dont une base de 100 000 ECG venant du Minnesota récupérée en 2015. Il faut payer pour, mais ce n'est pas trop cher ! Leur système est capable de prédire une centaine de troubles sur 12 canaux (ECG au repos) et une quinzaine sur 1 à 2 canaux (ECG ambulatoire). Le système détecte notamment la fibrillation atriale, qui est corrélée à l'apparition d'AVC – accidents vasculaires cérébraux – du fait d'une mauvaise circulation du sang, dont le risque augmente avec l'âge et qui est plus facile à traiter lorsqu'elle est précoce.

[528] Arterys a aussi obtenu en février 2018 l'agrément FDA pour la commercialisation d'une solution à base d'IA de détection de cancers du foie et du poumon. Voir Arterys Receives First FDA Clearance for Broad Oncology Imaging Suite with Deep Learning, février 2018.

[529] Le principe de la phase tomographie est expliqué dans Can Phase Space Tomography be Spun-off back into Medical Imaging? (10 pages).

L'analyse d'ECG passe souvent par l'utilisation des plusieurs méthodes et différents réseaux de neurones (récurrents, convolutionnels, autres)[530]. L'efficacité des méthodes dépend de la pathologie à détecter. Par exemple, la fibrillation attriale est mieux détectée avec un réseau convolutif. Cardiologs a obtenu l'agrément de la FDA en 2017 pour couvrir le marché US ce qui est assez rare pour une medtech française.

ECG en pratique…

les ECG sont analysés avec plusieurs méthodes :
CNN, RNN, FCNN
ANN, PNN
MLPNN, RBFNN, FFANN
LVQ, CBR
SOM-NN, F2FCM

spécialisées selon les besoins : métriques clés, arythmie, fibrillation atriale, etc.

ANN METHODS FOR ECG CLASSIFICATIONS

Author(s)	Feature extraction/reduction method	Classification model	Accuracy (%)
E. Derya et al. [19]	Eigenvector methods	RNN* MLPNN	98.06 for RNN 90.83 for MLPNN
I. Güler et al. [20]	DWT	ANN	96.94
S-N. Yu et al [21]	DWT	PNN	99.65
M. BEN MESSAOUD et al. [22]	The rate of heartbeat	RBFNN MLPNN	96 for MLP 85.2 for RBF
S. Meghriche, et al. [23]	Amplitudes, durations, and intervals of QRS, R, PP, RR, PR, and P waves.	CNN**	87.9
N. Belgacem et al. [24]	QRS complexes	MLPNN LVQ***	91.55 for MLP 89 for LVQ
G. K. Prasad et al. [25]	DWT	ANN	96.77
K. Lewenstein et al. [26]	Slope of an ST segment	RBFNN	97
N. Ouyang et al. [27]	Voltages of Q-, R-, S-, T-waveforms	FFANN	90.2 with anterior wall myocardial infarction (AI) 93.3 Without infarction.
Mozhiwen et al. [28]	Wavelet Transform	RBFNN	100 for trained samples 86.6 untrained samples
A. Rakotomamonjy [29]	DWT	ANN	79
B. Anuradha et al. [30]	Spectral entropy,	ANN	90
E. A. Fernandez et al. [31]	Attributes of the ECG	SOM-NN	90
U. R. Acharyaa et al. [17, 18]	Parameters that extracted from the ECG	ANN	85-100

*RNN: Recurrent Neural Network; **CNN: Compound Neural Network; ***LVQ: Learning Vector Quantization

HYBRID METHODS FOR ECG CLASSIFICATIONS

Author(s)	Feature extraction/reduction method	Classification model	Accuracy (%)
Y. zbay et al. [37]	Segments of arrhythmia.	MLP-BB+FCNN*	98.9 for ANN 99.9for FCNN
A. Sengur et al.[38]	Wavelet transforms and short time Fourier transform	AIS based fuzzy k-NN**	95.9 sensitivity# 96specificity# #
Z. Dokur et al. [39]	Fourier and wavelet analyses	ANN+GAs	96
K. Lewenstein et al. [40]	Segment of QRS complex, P and T wave	ANN + Expert System	92.5 sensitivity 96.7specificity
C-W. CHU et al. [41]	Moving average and differential equation approach	ANN and CBR	very high clustering performance
R. Ceylan et al. [42]	Segments of arrhythmia	T2FCM+ANN***	99
R. U. Acharya et al. [17, 18]	Spectral entropy	ANN + Fuzzy	80-85

*FCNN: fuzzy clustering neural network; **AIS: Artificial Immune System; ***T2FCM: Type Two Fuzzy C-mean. # Sensitivity: (true positive fraction) is the probability that a diagnostic test is positive, given that the person has the disease [4]. # # Specificity: (true negative fraction) is the probability that a diagnostic test is negative, given that the person does not have the disease [4].

A noter au passage que l'iWatch 4 annoncée par **Apple** en septembre 2018 intègre aussi une fonction d'électrocardiogramme. C'est un ECG à deux points qui ne détecte visiblement donc qu'environ 10% des pathologies cardiaques[531].

Il faut donc éviter de la survendre et ne pas oublier qu'un ECG réalisé chez un cardiologue sera plus riche en information ! Tout dépend de la pathologie recherchée ou à surveiller !

La prudence est autant de mise avec cette performance de **Google** et **Stanford** consistant à prédire des pathologies cardiaques par l'analyse du fond de l'œil. Réalisée en 2018, elle s'appuyait sur un jeu d'entraînement de 284 335 patients validé sur deux jeux de test de 12 026 et 999 autres patients. Le taux d'erreur semble rester élevé[532], et les chercheurs se gardent bien de le comparer avec celui des méthodes traditionnelles. Si le taux d'erreur d'un fond de l'œil est plus élevé qu'une échographie ou une IRM, pourquoi s'embêter à faire un fond de l'œil ?

Cancérologie

La cancérologie est l'un des plus gros marchés de la santé. Il est fragmenté en de nombreux types de cancers différents qui ont leurs propres techniques de diagnostics, s'appuyant sur différentes formes d'imagerie médicale et d'analyses biologiques. Presque tous les cancers ont leur solution spécifique d'analyse d'imagerie médicale à base d'IA.

Les **cancers du poumon** sont qualifiés par analyses de radios chez **Enlitic** (2014, USA, $15M) ainsi que chez **Riverain Technologies** (2004, USA) qui est focalisé sur la cage thoracique.

Le cancer du foie est aussi au programme. **Guerbet** (France, 467 M€ de CA en 2017) et IBM annonçaient en juillet 2018 le codéveloppement d'une solution dénommée Watson Imaging Care Advisor for Liver pour son diagnostic. Elle doit exploiter les résultats d'imagerie scanner et IRM et améliorer la caractérisation des métastases qui sont complexes à analyser dans ce type de contraste.

Infervision (2015, Chine, $75M) fait de la détection du cancer du poumon avec des scans PET. Le produit est déjà déployé dans divers hôpitaux en Chine. D'autres méthodes permettent cette détection mais avec des images 3D[533].

---

[530] Voir Machine Learning in Electrocardiogram Diagnosis, 2009.

[531] L'une des méthodes est décrite dans End-to-end Deep Learning from Raw Sensor Data Atrial Fibrillation Detection using Wearables 2018 (7 pages). Il s'agit ici de détecter la fibrillation atriale avec un réseau convolutionnel couplé à un LSTM (réseau à mémoire). Cette pathologie peut notamment entraîner un AVC.

[532] Voir Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning, février 2018 (21 pages).

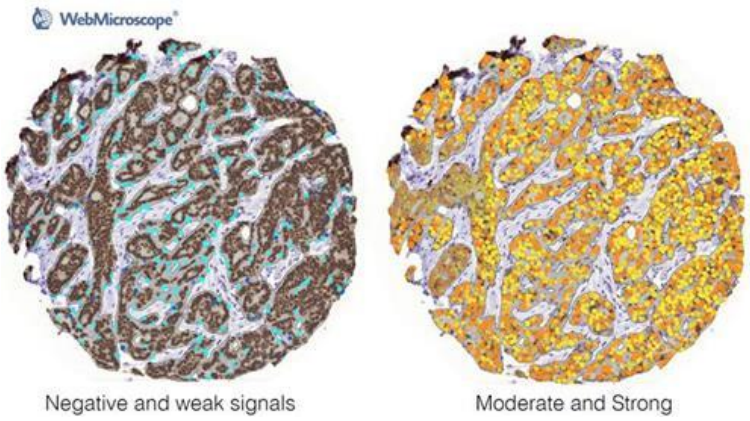
[533] Voir S4ND: Single-Shot Single-Scale Lung Nodule Detection, juin 2018 (8 pages).

La détection et qualification des **cancers du cerveau** sont aussi au programme des sances[534] ! Une startup francilienne, **Qynapse** (2015, France), analyse de manière itérative les résultats d'IRM cérébrales pour suivre l'évolution de traitements, notamment dans la lutte contre les cancers du cerveau.

Les **cancers du sein** sont analysés chez **Volpara Solution** (2009, Canada, $5,5M + IPO) qui réalise des analyses densitométriques précises et aussi chez **QVViewMedical** (2006, USA, $4,75M) ainsi qu'avec **Therapixel** (2013, France, $600K). Ces systèmes peuvent notamment être entraînés avec les 640 000 images issues de 86 000 patientes, récupérées dans la base du Digital Mammography DREAM Challenge lancé aux USA entre 2016 et 2017[535]. Therapixel avait d'ailleurs gagné mi 2017 la première étape de ce challenge. Therapixel se distingue des autres solutions en expliquant son diagnostic. Il se positionne en complément des spécialistes, pas pour les remplacer.

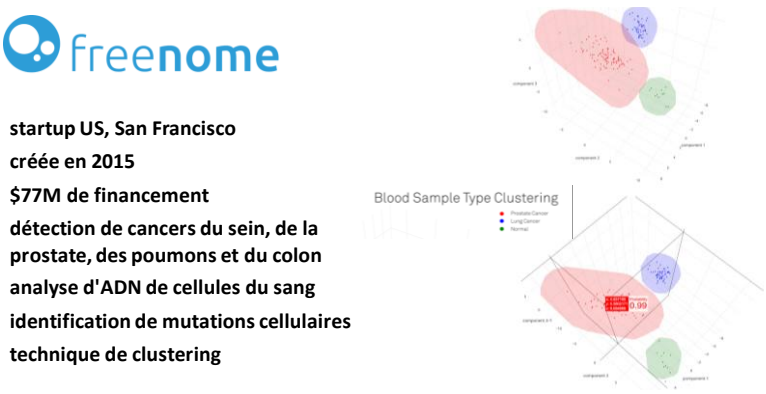
Imagia (2015, Canada) propose sa solution « Deep Radiomics » qui analyse l'évolution dans le temps de cancers par imagerie médicale couplée aux données cliniques des patients avec notamment le cancer du colon.

L'analyse de cellules cancéreuses dans des **biopsies** sur lamelles numérisées est proposée par le **WebMicroscope** (2013, Finlande, $6,5M, aussi dénommé Fimmic) qui réalise ses analyses dans le cloud à l'aide de GPU. Comme de nombreuses solutions d'IA en imagerie médicale, elle détecte des cellules cancéreuses et apporte surtout des indications quantitatives par comptage de cellules (*exemple ci-contre*).



KeeLab (2018, France) est positionné sur ce même marché de l'analyse en anatomie pathologique de cellules cancéreuses mais plus au niveau qualitatif. Ils cherchent à minimiser le taux de faux positifs, en-dessous de 1% sachant que c'est un outil d'aide au diagnostic pour les anapats. Ils ciblent les cancers du sein, du système digestif, de la prostate et du poumon. De son côté, **VitaDX** (2015, France) détecte le cancer de la vessie dans les cellules des urines avec un procédé voisin.

Freenome (2015, USA, $77,6M) produit des analyses de biopsies liquides – essentiellement du sang - permettant la détection de cancers émergents, du sein, de la prostate, des poumons et du colon (vidéo). Elle s'appuie sur des analyses génotypiques et sur le clustering de résultats. Reste après à traiter les cancers découverts !



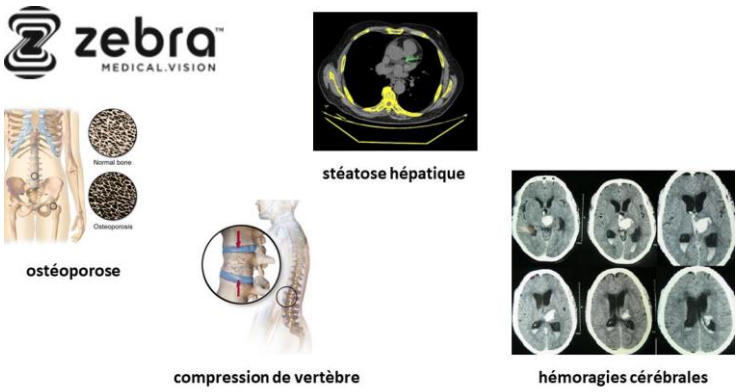
[534] Voir Brain Tumor Segmentation and Tractographic Feature Extraction from Structural MR Images for Overall Survival Prediction, 2018 (12 pages) appliquée à la détection précise de glioblastome, la pathologie qui a récemment emporté le sénateur américain John McCain.

[535] Voir https://www.synapse.org/#!Synapse:syn4224222/wiki/401743.

Squelette

Il fait l'objet de radios qui ne sont pas forcément difficiles à interpréter.

L'**ostéoporose** est pourtant caractérisée dans l'analyse de radios chez **Zebra Medical Vision** (2014, Israël, $50M) qui détecte aussi les compressions de vertèbres, la stéatose hépatique (foie gras) et les hémoragies cérébrales. Et ils détectent aussi le cancer du sein.



Système nerveux

Les **pathologies du cerveau** sont détectées par le Belge **Icometrix** (2011, $2,17M) avec son logiciel Msmetrix qui analyse les résultats d'imagerie médicale pour détecter les lésions, mesurer leur volume ainsi que celui du cerveau avec des applications dédiées à la sclérose en plaques. Des chercheurs anglais ont de leur côté créé une méthode de détection de lésions de petits vaisseaux dans le cerveau[536].

Max-Q.ai (2013, Israël, $2M) analyse les effets des accidents vasculaires cérébraux par l'analyse d'IRM en s'appuyant sur les briques de reconnaissance d'image d'IBM Watson, et semble-t-il, une une autre brique de détourage issue de **Mirada**. **Viz.AI** (2016, USA, $30,6M) a de son côté obtenu l'agrément FDA en février 2018 pour sa solution Contact qui réalise le diagnostic de l'artériosclérose cérébrale (vidéo).

Quantib (2012, Pays-Bas, 5M€) propose Quantib Neurodegenerative, une solution à base de deep learning qui quantifie et segmente les IRM du cerveau.

Diabète

L'IA intervient dans la gestion du diabète mais assez peu dans son diagnostic. La mesure de glycémie en continu est maintenant courante comme avec le patch FreeStyle d'**Abbot**. Ses résultats peuvent être exploités pour aider les patients à bien équilibrer leur diabète.

Au CES 2016, IBM présentait avec l'équipementier médical **Medtronic** une autre solution utilisant Watson pour prédire la survenue d'hypoglycémies des diabétiques de type 1. Les données exploitées étaient visiblement moins massives que celles de l'application sur les cancers.

L'hypoglycémie est générée par une boucle de rétro-action plus simple qui associe l'activité physique, la prise d'insuline et l'alimentation. Il faut donc mesurer les trois ce qui n'est pas trop compliqué pour les deux première mais moins évidente pour la dernière, même avec les capteurs de type Scio. Cependant, l'application est probablement pertinente pour ceux des diabétiques qui pratiquent un sport intensif et pour lesquels les risques d'hypoglycémie sont importants et répétés.

L'IA peut aussi intervenir dans la mesure de la glycémie. C'est le cas dans le prototype de « radar » la captant codéveloppé par l'**Université de Waterloo** au Canada, Google avec son kit radar Soli et l'Allemand Infineon. Le radar utilise la bande de 60 GHz, dans la tranche haute des ondes millimétriques (la 4G va utiliser le 28 GHz). L'outil sert surtout à la détection de mouvements de proximité comme avec un LeapMotion qui fonctionne dans l'infrarouge. Avec un peu de traitement du signal

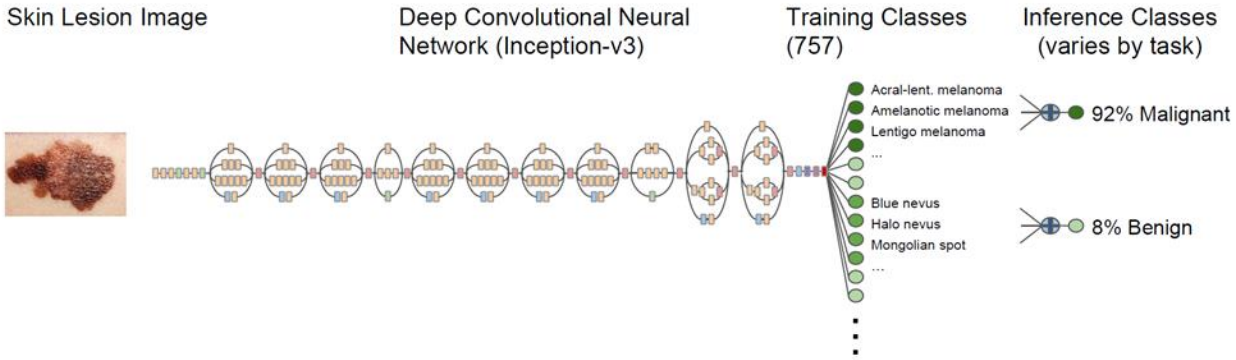
[536] Voir AI Detects and Measures Small Vessel Disease in Brain CT Scans, mai 2018.

et d'IA, les chercheurs ont réussi à mesurer le niveau de glycémie de manière totalement non invasive. Bon, pour l'instant, cela fonctionne en éprouvette mais pas sur de vrais gens [537]!

Dermatologie

L'imagerie médicale peut également jouer un rôle en dermatologie pour rendre cette discipline plus accessible, et pourquoi pas pour certains médecins généralistes.

Andre Esteva et Sebastian Thrun de Stanford décrivent très bien les enjeux et les méthodes utilisées dans la présentation Dermatologist-level classification of skin cancer with deep neural networks, 2017 (48 slides). La classification des lésions utilise une méthode assez classique avec un R-CNN pour les isoler puis un réseau convolutionnel, ici inception-v3[538], pour détecter les formes et un classifieur permettant d'isoler 757 variantes de cas dont 92% sont des lésions malignes (*ci-dessous*).



Une expérience concluante a été publiée en 2018 dans Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, 2018 par une équipe de chercheurs allemands (Heidelberg, Göttingen, Konstanz, Passau), américaine (New York) et française (Centre de Recherche en Cancérologie de Lyon) et relatée dans Artificial intelligence for melanoma diagnosis: how can we deliver on the promise?, août 2018.

Leur IA à base de deep learning génère un meilleur diagnostic de cancers de la peau que 58 dermatologues de référence. Le système avait été entraîné avec 100 000 images labellisées. Le taux de bonne reconnaissance était de 95% pour l'IA pour une moyenne de 89% pour les dermatologues.

Imagerie généraliste

De nombreuses startups se positionnent comme des généralistes couvrant plusieurs pathologies.

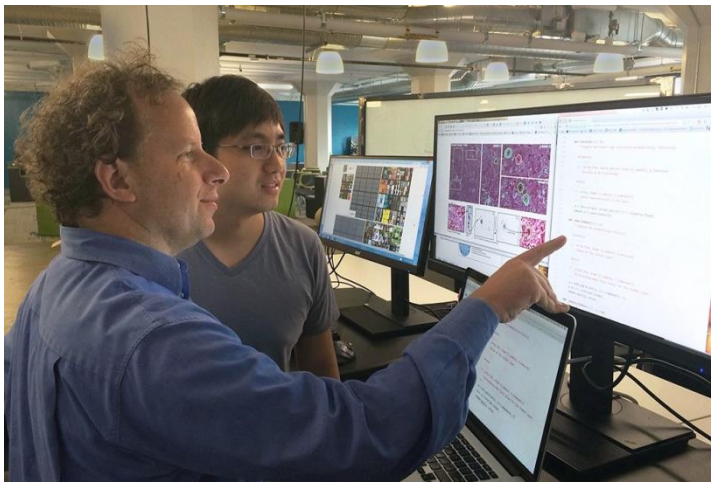
Cela peut sembler bizarre mais comme les algorithmes de détection sont assez voisins d'une pathologie à l'autre, pourquoi pas. Après, les modes de commercialisation ne sont pas les mêmes d'une spécialité à l'autre.

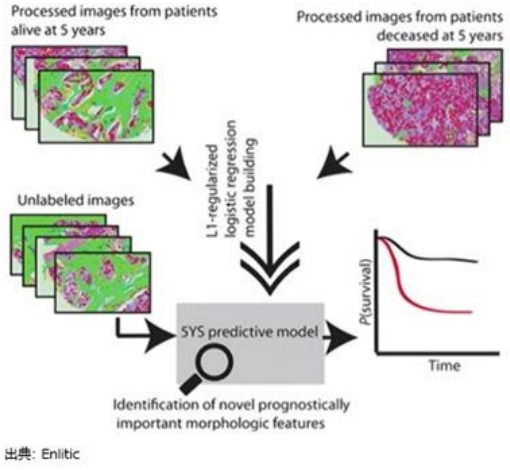
- **Enlitic** (2014, USA, $15M) qui propose de l'aide au diagnostic en s'appuyant sur les résultats de divers systèmes d'imagerie médical (IRM, scanner, radios) et sur du deep learning (*ci-dessous à gauche* avec son fondateur Jeremy Howard). Il détecte des pathologies émergentes le plus tôt possible, notamment les cancers du poumon. Il aide aussi à identifier plusieurs pathologies simultanément[539].

[537] Voir Artificial Intelligence and Radar Technologies to Measure Blood Glucose, juin 2018.

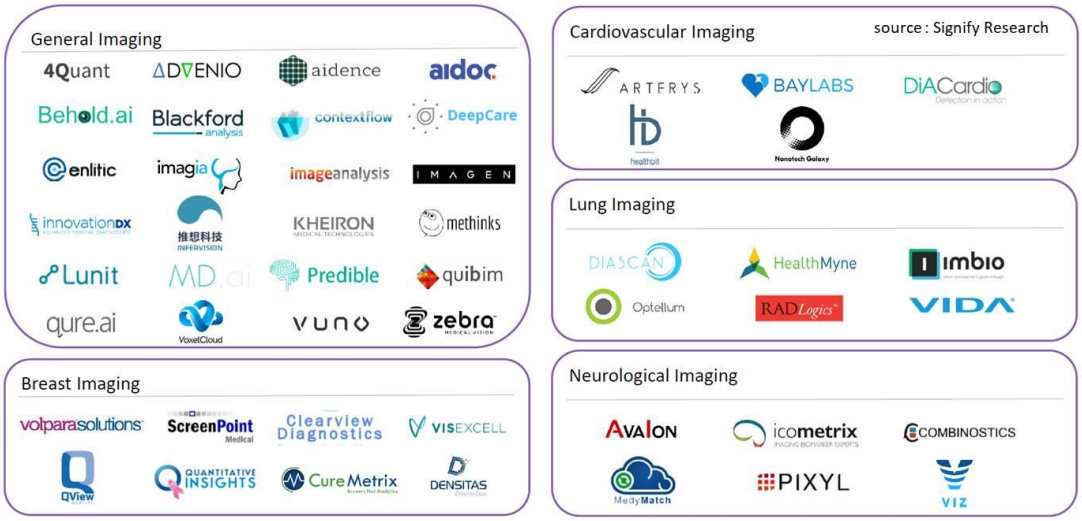
[538] Voir Rethinking the Inception Architecture for Computer Vision, d'une équipe de Google, 2015 (10 pages). Ce modèle de réseau convolutionnel présente l'avantage d'être peu couteux en temps machine pour son entraînement.

[539] Voir la **vidéo** de son CEO, Jeremy Howard à TEDx Bruxelles en décembre 2014. Il y aborde un point clé : il n'y a pas assez de médecins dans le monde. L'automatisation des diagnostics est donc un impératif incontournable.





- **VoxelCloud** (2015, USA, $28,5M) couvre le cancer du poumon, la rétinopathie diabétique, les maladies coronariennes et du foie. Bref, un peu de tout.
- **Lunit** (2013, Corée du Sud, $20.5M) propose une solution logicielle de deep learning générique d'interprétation d'imagerie médicale, notamment de radios et qui semble commercialisée en OEM et focalisée sur les poumons.
- **SigTuple** (2015, Inde, $25M) fait de l'analyse d'échantillons sanguins à base d'imagerie avec sa solution Shonit (vidéo) qui fait de comptage de cellules. Cela peut notamment servir à diagnostiquer l'anémie et la malaria.
- **Butterfly Network** (2011, USA, $100M) créé un échographe dont toute l'électronique tient sur un seul composant, et dont les images sont analysées par « computer vision ». Il semble qu'il utilise plusieurs émetteurs ultra-sons, un peu comme le système de OpnWatr qui fonctionne dans l'infrarouge pour faire de l'imagerie cérébrale. Son fondateur Jonathan Rothberg a de l'expérience, ayant créé et revendu deux sociétés de séquençage de l'ADN, 454 à Roche et Ion Torrent à Thermo Fisher.
- Et il faut évidemment compter avec **IBM** qui développe une panoplie de solutions d'interprétation d'imagerie médicale avec Watson Health et dont les solutions logicielles sont aussi exploitées par bon nombre de startups du secteur[540].

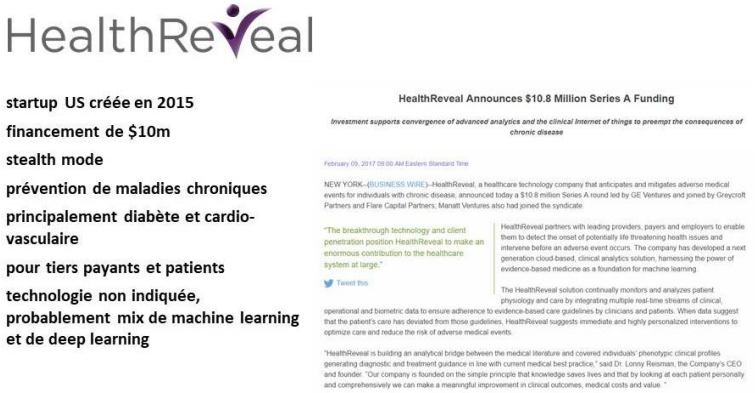


[540] Voir cette intéressante analyse de la position d'IBM Watson dans l'imagerie médicale : http://www.nanalyze.com/2017/08/ibm-dominate-radiology-ai/ ainsi que IBM's Automated Radiologist Can Read Images and Medical Records de Tom Simonite, février 2016 dans la MIT Technology Review.

L'imagerie médicale n'est pas la seule source de diagnostics médicaux. Il faut ajouter l'analyse d'EEG (électroencéphalogrammes), les tests biologiques (sang, urine) ainsi que les tests d'ADN (génotypie et séquençage complet).

De nombreuses startups ambitionnent d'exploiter tout ou partie de ces données pour améliorer les diagnostics, surtout dans le cadre de médecine préventive et pas seulement curative.

HealthReveal (2015, USA, $11,3M) propose une solution en cloud de prévention de l'apparition de maladies chroniques liées au style de vie, basée sur l'utilisation de capteurs biométriques divers. La solution en cloud exploite ces données en plus des données de parcours de santé. es clients sont plutôt les tiers payants dans la santé, surtout sur le marché des USA.



Forward (2016, USA) est une étonnante startup qui veut réinventer le cabinet médical. Son premier site à San Francisco est équipé de tous les capteurs[541], outils d'analyses de laboratoires, ADN compris, et systèmes d'imagerie médicale pour faire un bilan de complet à 360°[542]. Ce n'est pas une clinique pour autant. L'offre qui est proposée sur abonnement se veut être une solution de médecine préventive.

Génomique

La génomique est un domaine plein de promesses pour ce qui est de la médecine autant prédictive que curative.

Nombre de startups se sont lancées dans la collecte du génotype ou du génome des patients, la différence entre les deux n'étant pas encore bien comprise par le grand public[543]. Les données collectées dans les deux cas sont de grande taille pour chaque patient.

Ce qui fait qu'en termes de « big data », la génomique et toutes les autres « omics » (proteomics, bacteromics, etc), les outils d'analyse font appel à des données différentes.

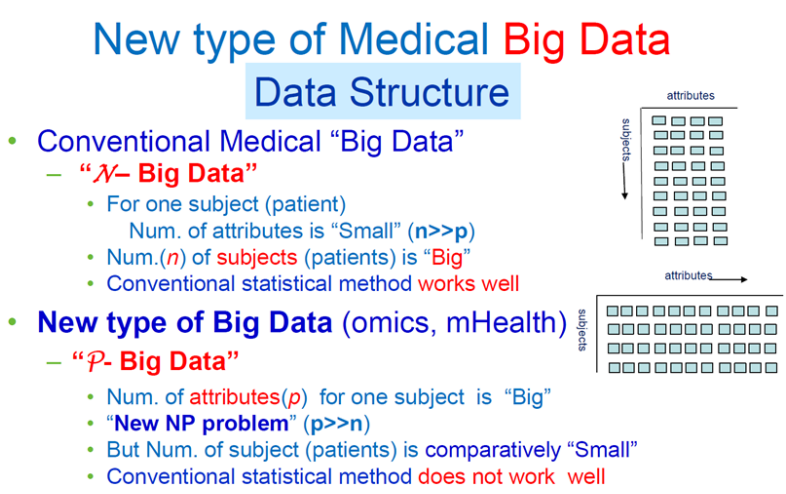
---

[541] Ils ont même développé leur propre scanner corporel qui capte à distance le pouls et la température corporelle. Mais cela ne remplace pas une IRM ou une tomodensitométrie.

[542] Loïc Le Meur a filmé avec son smartphone une visite assez complète de Forward à San Francisco et c'est très instructif : https://www.facebook.com/loic/videos/381807855521818/.

[543] Le génotype proposé par exemple par 23andme consiste à évaluer les différences entre le génome standard humain et votre génome. Environ 500 000 différences sont évaluables par les techniques courantes de génotypage, qui ont un prix public d'environ $100. Le séquençage du génome complet consiste à analyser en détail les trois milliards de bases de l'ensemble de nos chromosomes, dont seulement 1,5% comprend la partie codante des gênes servant à générer des protéines. En gros, le séquençage génère beaucoup plus de données que le génotypage. Mais on ne sait pas encore très bien l'exploiter pour faire de la médecine prédictive et curative.

Dans le big data traditionnel, on a un nombre limité de paramètres pour un grand nombre d'engistrements, comme les caractéristiques de clients dans une base de clients. Dans les « omics », on a un très grand nombre de paramètres, qui s'évaluent en centaines de milliers à milliards, et avec un nombre généralement plus limité d'enregistrements (patients). Les plus grosses bases de génomes en comprennent aujourd'hui quelques dizaines de milliers[544].



L'IA et le machine learning sont là pour exploiter tout cela !

Portable Genomics (2011, USA) est une startup créée par des français implantés à San Diego aux USA. Elle a créé une solution logicielle mobile de collecte et de visualisation des données de santé d'une personne. Elle en assure le stockage en ligne, sous le contrôle de l'utilisateur.

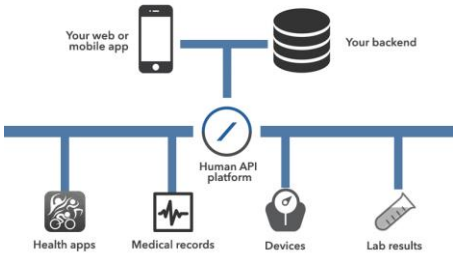
La solution collecte aussi bien les données de génomique issues d'un séquençage (complet du génome) ou d'un génotypage (analyse de variations types des gênes) que celles de la santé en général : historique des pathologies, mode de vie et données issues d'objets connectés. Cela permet de constituer une vue à 360° du patient, utile pour les praticiens.

Human API PORTABLE GENOMICS

deux startups US qui consolident les données de santé à grande échelle

les commercialisent ensuite sous contrôle des patients aux professionnels de santé

permettent des analyses diverses à base d'IA



Ca l'est aussi pour créer des bases de données santé exploitables par la recherche et les entreprises de pharmacie. Cela permet aussi d'identifier le niveau de risques de diverses pathologies. La société se positionne comme une plateforme de collecte, de partage et de monétisation de données personnelles de santé, s'appuyant sur un modèle de partage de revenu avec les utilisateurs.

Deep Genomics (2014, USA/Canada, $16,7M) a créé le DG Engine qui analyse les variations du génome – les mutations de l'ADN – et la manière dont elles affectent le fonctionnement des cellules et génèrent des pathologies.

Ce sont des "genome-wide association study" (GWAS) qui produisent des analyses de corrélations entre modifications des gênes et pathologies (le "phénotype"). Les analyses réalisées par Deep Genomics ont la particularité d'intégrer tout le cycle de vie des gênes et notamment leur épissage – qui correspond à l'extraction de la partie codante des gênes – jusqu'à leur translation, à savoir la conversion de l'ARN qui résulte de l'épissage en protéines dans les ribosomes.

Ils proposent en open source leur base de données SPIDEX de mutations de gênes et de leurs effets sur leur épissage[545].

---

[544] Le schéma vient de la présentation Big data and artificial intelligence in medicine and drug discovery de Hiroshi Tanaka (39 slides).

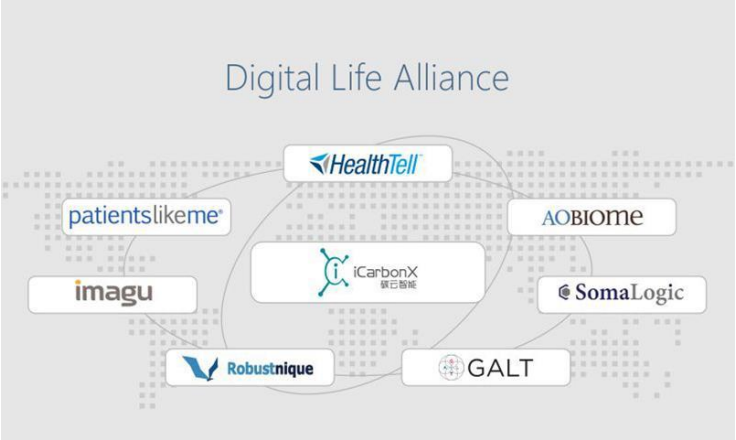
[545] Voir The human splicing code reveals new insights into the genetic determinants of disease, 2015, qui explique les fondements scientifiques de leur procédé.

L'ambition est de mener à de la médecine personnalisée mais on en est encore loin. La société a été cofondée par Brendan Frey, qui avait fait son PhD à Toronton avec Geoff Hinton, un chercheur canadien à l'origine du décolage du deep learning en 2006 et qui est maintenant chez Google.

Pathway Genomics (2008, USA, $43M) est une société américaine qui propose divers tests génétiques et biopsies ciblés par risques pathologiques permettant d'identifier des facteurs de risques divers et variés en cardologie (Cardia DNA Insight), dermatologie (SkinFit), BRCATrue (cancer du sein), ColoTrue (cancer du colon) et obsésité (Healthy Weight DNA Insight). Le test Mental Health DNA Insight permet d'évaluer l'impact des traitements en psychothérapies et le Pain Medication DNA Insight évalue l'efficacité probable des analgésiques. La société utilise IBM Watson. Ici, on a surtout affaire à un bon packaging par pathologie car les données exploitées par ces différents tests proviennent généralement des mêmes analyses, comme la génotypie réalisée par 23andme qui analyse plus de 500 000 variations dans les gênes (génotypie à base de SNP, ou single nucleotid polymorphisms).

iCarbonX (2015, Chine, $200M, ils annoncent $600M, mais dont $400M correspondent à des investissements chez leurs partenaires[546]) veut faire des données médicales, génomiques comprises, une plateforme pour des applications en propre et tierce-partie. La société collecte les données de santé de patients à 360° : génomique, phénotype, bactérome, protéome, données comportementales et psychologiques (on est en Chine…). Le tout exploite de l'IA dont la nature n'est aucunement précisée, tout comme n'est pas détaillée l'origine des données exploitée. Comment les patients fournissent-ils leurs données à iCarbonX ? Ils ont déjà une demi-douzaine de partenaires applicatifs dans leur Digital Life Alliance qui exploitent leurs données.

Avec SomaLogic (1999, USA, $395M, analyse biologique de protéines), HealthTell (2010, USA, $40M, évaluation de la réaction immunologique), PatientsLikeMe (2004, USA, $127M, réseau social de santé), AOBiome (2013, USA, $38,7M, thérapies probiotiques), GALT (General Automation Lab Technologies, 2014, USA, $10,5M, analyses du microbiome) et Robustnique (2010, Chine, cosmétique à base d'enzymes).



Le dernier est Imagu Vision Technologie (2005, Israël) et a été acquis en 2016. Ils apportent l'analyse d'imagerie médicale à iCarbonX, permettant de compléter le profil santé des patients. Ce qui leur a permis, ipso-facto, d'établir un laboratoire de R&D en Israël.

Sophia Genetics (2011, Suisse, $63M) propose une solution de diagnostic basée sur l'analyse du génôme. Elle déjà déployée dans plusieurs centaines d'hôpitaux dans le monde.

Ginger.io (2011, USA, $28,2M) a créé un outil de diagnostic et de prescription de traitement pour diverses pathologies neuropsychologiques. Il exploite des applications mobiles pour le diagnostic et du machine learning. La solution permet un auto-traitement de certaines pathologies par les patients.

Lumiata (2013, USA, $31M) est dans la même lignée un système d'analyse de situation de patient permettant d'accélérer les diagnostics, notamment en milieu hospitalier.

[546] Voir Chinese AI company plans to mine health data faster than rivals, dans Nature, janvier 2017.

MedWhat (2010, USA, $3,2M) propose une solution générique d'aide au diagnostic qui s'appuie sur la panoplie totale de l'IA (deep learning, machine learning, NLP). Elle se matérialise sous la forme d'une application mobile faisant tourner un agent conversationnel à qui ont indique ses symptômes, qui pose des questions de qualification et oriente ensuite le patient (**vidéo de démo**). Elle stocke aussi le dossier médical du patient. La startup a été créée par des anciens de Stanford, mais cela ne semble pas suffisant pour décoller !

Dans les applications santé d'IBM Watson, on peut citer l'application de **GenieMD** (2010, USA, $100K) qui permet aux patients, aux USA, de faire un premier niveau d'autodiagnostic de problèmes de santé courants et d'être ensuite mis en relation avec des praticiens. Il permet aussi de suivre l'observance de la prise de médicaments. La solution exploite les informations fournies par les patients en langage naturel. C'est une application générique qui pourrait être mise en oeuvre dans les stations de télémédecine pour les déserts médicaux.

Télémédecine

Babylon Health (2013, UK, $85M) est une startup qui propose la combinaison de l'accès à un chatbot médical pour le diagnostic de premier niveau puis à un docteur en ligne pour poursuivre la discussion dans une visioconférence. Le diagnostic de premier niveau serait meilleur que celui des généralistes débutants au Royaume Uni (81% vs 72%).

Le chatbot s'appuie sur un moteur de NLP pour comprendre les questions du patient et dialoguer avec lui, un Knowledge Graph pour modéliser les connaissances, un moteur d'inférence (système expert) pour réaliser son diagnostic, sachant qu'il est combiné à des modèles probabilistes et du machine learning.

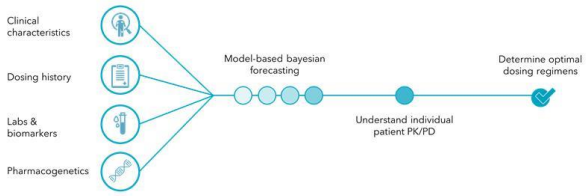
En Chine, le robot médecin **Xiaoyi** serait capable depuis fin 2017 d'égaler les médecins généralistes dans le pré-diagnostic, surtout en hôpital. Il aiderait les praticiens à remplir plus rapidement les dossiers des patients. Mais l'histoire ne dit pas à quelles informations des patients il a accès (source).

Médecine prédictive généraliste

Prognos (2010, USA, $43,2M) est l'un des acteurs clés qui se positionnent dans le champ de la médecine prédictive. Ils ambitionnent de détecter le plus en amont possible une cinquantaine de pathologies en s'appuyant sur une gigantesque base de données de plus de 13 milliards de résultats d'examens biologiques de plus de 160 millions de patients. Le tout doit s'appuyer sur force machine learning avec analyses de corrélation entre paramètres biologiques et occurrences de pathologies. La startup annonce avoir créé plus de 1000 algorithmes propriétaires, et visiblement, sans données génomiques.

Medopad (2011, UK, $32M) adopte une approche voisine en collectant un maximum d'informations issues d'objets connectés comme les montres intégrant un capteur cardiaque, les trackers d'activité et autres lecteurs de glycémie ou de tension pour les analyser et détecter des états de santé nécessitant un traitement. Leur solution est commercialisée tout azimut, y compris aux assurances santé privées qui peuvent mettre en place des bonus/malus en fonction du comportement des assurés.

InsightRX (2015, USA, $2,8M) utilise le machine learning et les données patients pour optimiser le choix des thérapies et leur dosage. Les données comprennent cette fois-ci les informations de génomique. La solution serait déjà déployée dans des hôpitaux aux USA.



Thérapies

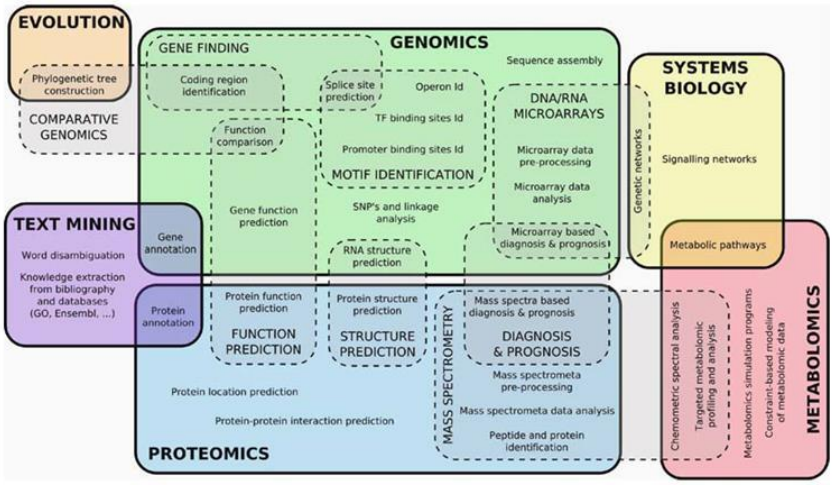
Passons maintenant à l'exploitation d'IA pour guérir !

Les biotechs sont de grandes consommatrices de logiciels et d'IA. En cause, les volumes de données à gérer et analyser, à commencer par ceux qui viennent de toutes les techniques en « omique » : la génomique (analyse de l'ADN et de l'ARN) et la protéomique (analyse des protéines). La baisse du cout du séquençage de génomes de toutes les espèces vivantes a généré d'énormes quantités de données à exploiter.

L'IA peut aider à comprendre la structure des gênes et de leur expression, l'épissage des gênes (comment les différentes parties d'un gêne s'assemblent), le repliement des protéines sur elles-mêmes après leur production dans les ribosomes des cellules ou la détermination des paramètres qui favorisent ou pas l'expression des gênes, notamment ceux qui déclenchent des cancers[547].

Le schéma ci-dessous illustre la variété des usages de l'IA dans ces domaines (source). Les techniques employées tournent essentiellement autour du machine learning. Le deep learning commence aussi à faire son apparition dans certains cas d'usages listés ci-dessous[548].

Le deep learning permet de réaliser des prévisions de comportements de molécules et de structures de protéines, des problèmes mathématiques très complexes à résoudre par des méthodes traditionnelles[549]. Et nous n'en sommes qu'au début dans ce domaine, l'informatique quantique pouvant à plus long terme permettre d'aller encore plus avant dans ces simulations.



deep learning

	Omics
	Research topics
Deep neural network	Protein structure
	Gene expression regulation
	Anomaly classification
Convolutional neural network	Gene expression regulation
Recurrent neural network	Protein structure
	Protein classification
	Gene expression regulation

[547] Voir Deep Learning in Pharmacogenomics From Gene Regulation to Patient Stratification 2017 (40 pages) qui fait un bon tour d'horizon de méthodes de deep learning utilisées dans la recherche en génomique.

[548] Voir Deep Learning in Bioinformatics des Coréens Seonwoo Min, Byunghan Lee et Sungroh Yoon, 2016 (40 pages).

[549] Voir Scientists develop machine-learning method to predict the behavior of molecules, octobre 2017.

L'IA ne permet cependant pas pour autant de créer cette fameuse « médecine de précision » 100% adaptée à chaque individu et en fonction de son génome. C'est pour l'instant bien trop complexe, ne serait-ce que d'un point de vue biologique[550].

Bien qu'utilisés essentiellement en imagerie médicale, les réseaux de neurones convolutionnels sont aussi exploités en génomique. Mais ce sont les réseaux de neurones récurrents (RNN) qui sont plus couramment employés, car ils sont adaptés à l'analyse de données séquentielles comme pour le langage, or l'ADN est un langage, à base de quatre lettres (ATCG).

Nous ne traiterons pas ici du sujet dans ses recoins mais allons plutôt l'illustrer par quelques startups actives dans le domaine comme presque partout ailleurs dans ce document.

Drug discovery

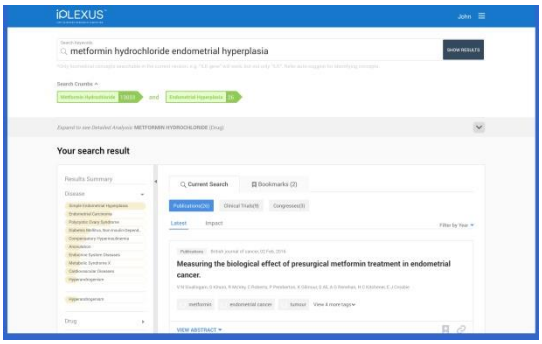
L'usage de l'IA dans la découverte de nouvelles molécules thérapeutiques est un champ d'exploration relativement nouveau et immature. Le machine learning peut aider à exploiter de gros volumes de données pour faire ressortir des molécules intéressantes, notamment par la méthode du criblage, qui associe des molécules connues à différentes cibles thérapeutiques.

D'autres essayent de modéliser la structure tridimentionnelle des molécules organiques (peptides, polypeptides, protéines, enzymes) pour créer ou identifier des zones actives. Cela fait partie du vaste champ de la simulation du repliement des protéines. Il existe des méthodes s'appuyant sur du crowdsourcing comme via le site FoldIt et d'autres qui exploitent du deep learning.

Il existe même des méthodes de création de nouvelles molécules thérapeutiques exploitant un réseau de neurones génératif comme ReLeaSE, créé par l'Université de Caroline du Nord[551]. Enfin, à plus long terme, on pourra peut-être le faire via des algorithmes quantiques sur des ordinateurs quantiques disposant d'un très grand nombre de qubits intriqués. Tout cela fait partie du vaste champ de la bioinformatique.

Voici quelques startups et projets de recherche qui s'activent dans le domaine de la découverte de nouvelles thérapies.

Innoplexus (2011, Allemagne/Inde) propose iPlexus, un moteur de recherche d'informations médicales qui exploite 27 millions de publications, 365 000 rapports de tests cliniques et un million de thèses. Le tout s'appuie sur du machine learning pour générer des analytics quantitatifs et du traitement du langage pour les recherches. La solution exploite aussi une Blockchain pour assurer la traçabilité des données intégrées dans la base. En plus de la pharmacie, la startup vise aussi les marchés financiers.



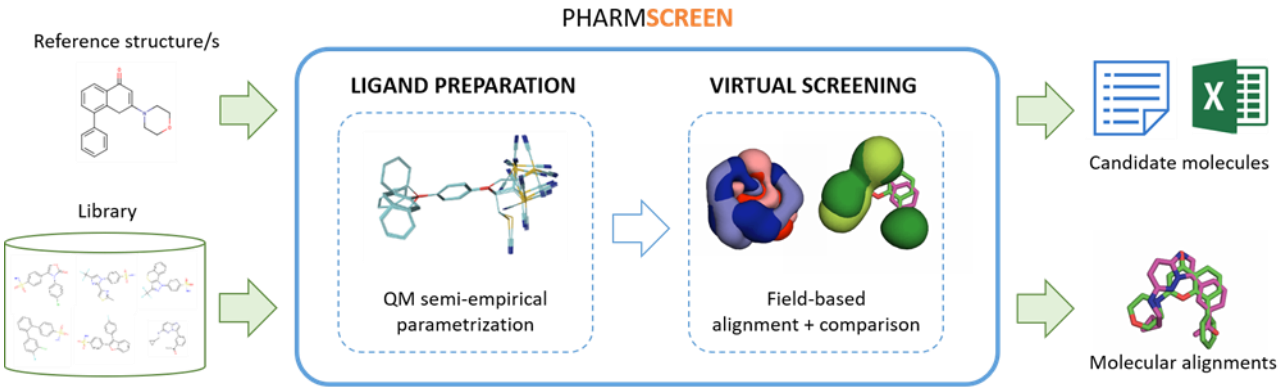
[550] Voir cet excellent papier qui remet les pendules à l'heure sur la médecine de précision : Why Does the Shift from "Personalized Medicine" to "Precision Health" and "Wellness Genomics" Matter? d'Eric Juengst et Michelle McGowan, septembre 2018.

[551] Voir Artificial intelligence system designs drugs from scratch, juillet 2018 et Deep reinforcement learning for de novo drug design, 2018 (27 pages).

IKTOS (2016, France) utilise du deep learning pour réaliser des simulations biologiques de l'effet de médicaments. L'idée consiste à screener des molécules existantes et à identifier in-silico leurs interactions avec des protéines connues selon un cahier des charges donné d'attaques de cibles à des fins thérapeutiques. Ils exploitent pour cela un réseau de neurones qui converti la structure des molécules connues dans un langage intermédiaire qui est ensuite rapproché des protéines cibles[552].

Cela fait penser à la notion de *stacked autoencoder*. Ils ne sont pas seuls sur ce marché qui comprend d'autres startups telles que Certara, ChemAxon, Mind the Byte, Optibrium et Tripos.

Pharmacelera (2015, Espagne, 290K€) a développé Pharmascreen, un moteur de recherche de molécules en exploitant des algorithmes de modélisation à base d'IA, no précisés.



Numerate (2007, USA, $17,4M) utilise l'IA pour aider à la découverte de nouveaux traitements à base de petites molécules pour traiter divers cancers, maladies neurodégératives et cardiaques (ce qui couvre un bel éventail des besoins du marché). La startup travaille avec Merck. **Exscientia** (2012, UK) semble positionné sur le même créneau.

Insilico Medicine (2014, USA, $20M) fait partie du grand nombre d'acteurs qui cherchent à trouver de nouvelles solutions curatives contre le cancer et les maladies du vieillissement à base de génomique et de big data. C'est en fait un prestataire de services qui créé de nombreuses solutions ad-hoc à base de deep learning. Il aide notamment d'autres entreprises à identifier de nouvelles thérapies, comme Pharmaceutical Artificial Intelligence. Leur logiciel en ligne aging.ai vous permet de déterminer votre âge à partir de vos résultats d'analyse sanguine. Mais vous pouvez aussi vous rappeler de votre date de naissance, ou au pire, consulter votre carte d'identité dans le pire des cas, ce qui sera probablement plus fiable.

Envisagenics (2014, USA, $4,6M) réalise des analyses d'épissage de l'ARN pour détecter les cancers provoqués par la malformation de protéines et faciliter le développement de nouvelles molécules thérapeutiques qui corrigent ces problèmes d'épissage (vidéo). L'IA exploitée dans ce genre de solution n'est pas précisée.

[552] Une méthode qui semble voisine a été conçue par une équipe du MIT pilotée par Wengong Jin en et Regina Barzilay en 2018. Elle s'appuie sur une modélisation des molécules organiques sous forme de graphes entraînée avec 300 000 molécules différentes connues. Le projet a éré mené dans le cadre du consorsium Machine Learning for Pharmaceutical Discovery and Synthesis piloté par le MIT et des entreprises de pharmacie, et financé en partie par la DARPA. Voir Junction Tree Variational Autoencoder for Molecular Graph Generation, 2018 (17 pages).

Atomwise (2012, USA, $51,3M) est une startup américaine qui utilise le machine learning pour découvrir de nouveaux médicaments et vérifier leur non toxicité. Le principe consiste à simuler l'interaction entre des milliers de médicaments connus et une pathologie telle qu'un virus, et d'identifier celles qui pourraient avoir un effet par simulation des interactions moléculaires. Un premier résultat aurait été obtenu en 2015 sur un virus d'Ebola. La simulation in-silico permet de choisir quelques médicaments qui sont ensuite testés in-vitro avec des cellules humaines. Merck fait aussi appel à Atomwise.

Owkin (2016, France/USA, $18,1M) est sur un créneau voisin que l'on appelle le « drug repositioning » qui permet d'étudier des essais cliniques et d'évaluer l'intérêt de certains médicaments sur d'autres pathologies que celles qui ont été testées. Le tout s'appuie sur du machine learning. Ils sont partenaires de l'INSERM.

Standigm (2015, Corée du Sud, $3,7M) utilise le machine learning pour réduire les risques d'échecs lors de la phase I de tests cliniques de nouvelles thérapies hybrides associant plusieurs thérapies existantes[553].

Dexstr.io (2014, France) est une startup toulousaine fournissant la solution Inquiro qui exploite les données médicales non structurées pour faciliter la recherche d'informations pour les sociétés de pharmacie.

En gros, c'est de la recherche documentaire, un peu comme le font Sinequa et Antidot, mais avec un tuning adapté à la documentation scientifique dans la santé. Leur concurrent serait plutôt l'application d'IBM Watson à l'oncologie.

Berg (2006, USA) utilise de l'IA pour faire de la recherche thérapeutique en biotechs. Leur IA croise des données biologiques et phénotypales (environnement, …) des patients. Ils exploitent diverses données : génomiques, protéomiques, lipidomiques, et métabolomiques, ainsi qu'au niveau du fonctionnement du cycle énergétique dans les mitochondries[554]. Leur système de simulation « in silico » permettrait de simuler l'efficacité de thérapies, notamment contre les cancers et les maladies neurodégénratives. Ils ont créé une molécule qui restaure le processus de l'apoptose (la mort programmée) de cellules cancéreuses (BPM 31510). Elle est en tests cliniques. Ils sont partenaires, entre autres, de Sanofi Pasteur en France pour développer un vaccin antigrippal fonctionnant d'une année sur l'autre ainsi qu'avec Astra Zeneca sur le traitement de la maladie de Parkinson.

On peut aussi prédire certaines contre-indications de combinaisons de médicaments, comme avec le système **Decagon** créé par Marinka Zitnik, Monica Agrawal et Jure Leskovec de l'Université de Stanford. La méthode utilise du deep learning de modélisation de graphes d'interactions entre molécules dans un réseau de neurones[555]. L'IA peut aussi servir à identifier la **toxicité** de certaines molécules et de réduire la quantité de gtests à réaliser sur des animaux dans les tests pré-cliniques[556].

Au passage, notez que je ne cite pas de grand laboratoire de pharmacies. Il se trouve qu'ils communiquent peu sur leurs avancées en matière d'usages de l'IA et sous-traitent une grande partie de leur innovation amont dans les biotechs. On n'en parle pas souvent mais ces entreprises sont menacées non pas par les GAFA mais par des portfeuille de brevets veillissants et par une position bancale dans la chaine de valeur. Cela se manifeste avec une rentabilité financière en baisse constante depuis une vingtaine d'années[557].

---

[553] Le procédé utilisé par Standigm est décrit ici : http://www.standigm.com/project/.

[554] Voir How artificial intelligence is changing drug discovery de Nic Fleming dans Nature, mai 2018.

[555] Voir Modeling polypharmacy side effects with graph convolutional networks, 2018 (9 pages et version en slides).

[556] Voir Database analysis more reliable than animal testing for toxic chemicals du Johns Hopkins University Bloomberg School of Public Health, 2018.

[557] Voir Pharma's broken business model: An industry on the brink of terminal decline, de Kevin Slott, novembre 2017 qui décrit très bien les soubressauts qui affectent l'activité et le business model des industries de la pharmacie.

Cancérologie

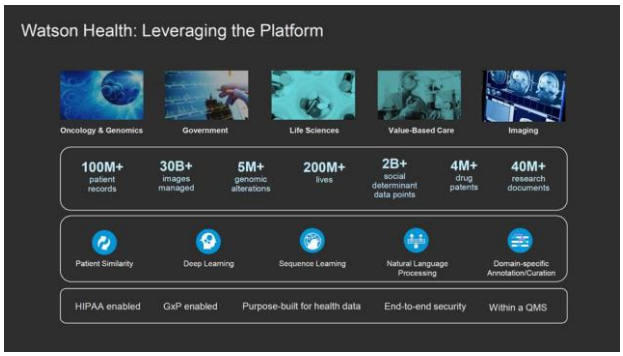
IBM a été incontestablement l'un des premiers grands acteurs à relancer la thématique de l'intelligence, bien avant que les GAFA s'en mêlent. La date clé à retenir est 2011 avec la victoire d'IBM Watson dans le jeu **Jeopardy** face aux meilleurs joueurs américains. Watson était au départ surtout un système de traitement du langage et de modélisation des connaissances. Aujourd'hui, c'est un ensemble de briques logicielles assemblées par les développeurs indépendants et par les équipes services d'IBM.

IBM Watson est décliné sur divers cas d'usages dans différents marchés verticaux. Il est important de comprendre que Watson n'est pas un produit plurivalent mais un ensemble de briques logicielles exploitées dans des solutions métiers développées par IBM ou ses partenaires. Il en va de même pour les briques logicielles d'IA, peut-être moins bien marketées, que l'on trouve chez Google, Microsoft, Amazon et d'autres grands du logiciel et de l'Internet.

L'un des premiers marchés qu'IBM a cherché à pénétrer est celui de la santé et en particulier celui de la cancérologie. Leur ambition était d'utiliser Watson à la fois pour aider au diagnostic puis à la recommandation de traitements, le tout réalisé en partenariat avec de nombreuses cliniques américaines et dans le monde, et en exploitant les données phénotypiques (présence de la maladie) et génotypiques du patient (variantes dans l'ADN des gênes) et les bases de connaissance du secteur composées de millions de documents (recherche, études cliniques).

La solution **Watson for Oncology** a été créé initialement en partenariat avec l'assureur santé Anthem (anciennement WellPoint) et le Memorial Sloan Kettering Cancer Center (MSK) de New York, qui associe un hôpital et un centre de recherche.


en cancérologie, avec Mayo Clinic, Wellpoint, ...

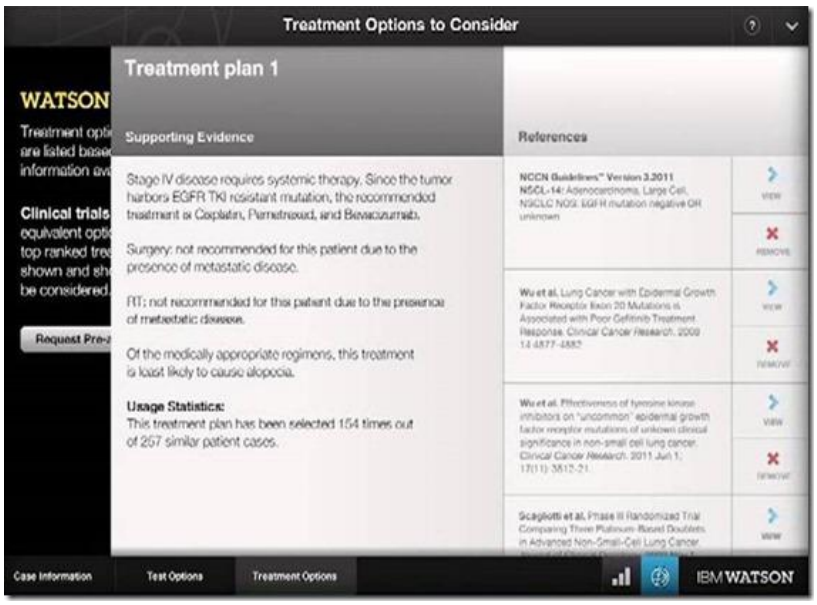


Elle a ensuite été déployée timidement dans plus d'une quinzaine d'établissements aux USA et ailleurs dans le monde comme en Inde, mais sans que l'on sache auprès de combien de praticiens et sur combien de cas de patients.

IBM Watson analyse les dossiers de patients atteints de tumeurs bénignes, y compris le séquençage d'ADN de cellules prélevées dans les tumeurs[558], propose un diagnostic, détermine des traitements possibles et évalue leur efficacité relative. Il aide notamment à optimiser l'usage de la chirurgie, de la radiothérapie et de la chimiothérapie. Les cancers sont des pathologies idéales pour Watson car elles sont plurifactorielles. Mais ce n'est pas (encore) de la médecine préventive ou prédictive.

[558] Semble-t-il, et non pas un simple génotypage, mais on peut aussi séquencer l'ARN qui évalue l'expression des gênes dans les tumeurs.

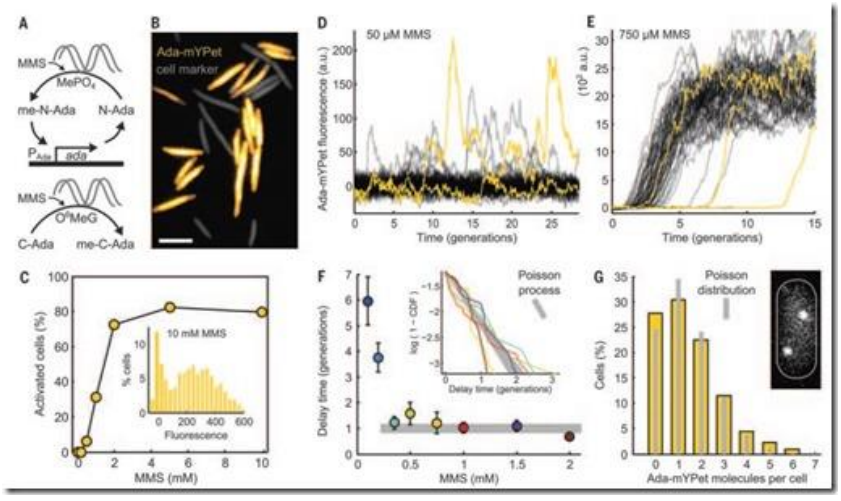
En 2014, le **Baylor College of Medicine** créait l'application KnIT (Knowledge Integration Toolkit) s'appuyant sur IBM Watson pour identifier des thérapies contre le cancer. Précisément, elle analysait la littérature scientifique pour suggérer six protéines kinases capables de contrôler le fonctionnement de la protéine p53 qui jouerait un rôle dans le développement d'environ la moitié des cancers. En 30 ans, selon IBM, moins d'une trentaine de nouvelles protéines auraient été découvertes. Ce qui mériterait d'être vérifié !



Des données statistiques peuvent exister qui font le lien entre type de thérapies et types de mutation de ces gênes. On est ici dans le domaine du big data non structuré contrairement au big data dans le marketing qui est basé sur des données bien plus structurées en général (logs Internet, données d'achats ou de consommation, bases de données relationnelles, etc). Il semble que cette partie de la solution ait été développée en partenariat avec **Cleveland Clinic**.

La solution utilise des sources d'informations variées pour faire son diagnostic et pioche notamment dans les 44 000 nouvelles publications scientifiques annuelles sur le cancer.

Les articles ne sont pas toujours faciles à exploiter : autant le texte est relativement facile à analyser, autant les illustrations qui ne sont pas toujours fournies dans un format structuré ne sont pas facilement exploitables.



Or elles fournissent des données critiques, exploitables statistiquement, à supposer que Watson puisse comprendre leur signification.

L'exploitation de la littérature scientifique ne doit donc pas être bien évidente à ce niveau. Par contre, elle est peut-être plus aisée pour les études liées aux AMM (autorisations de mise sur le marché) et autres études épidémiologiques. On se demande par contre si ce genre de solution sait tenir compte de la forte proportion de publications scientifiques qui est entachée de fraudes ou exagérations[559].

Dans les démonstrations d'IBM, la solution de cancérologie à base de Watson fournit au praticien un choix de traitements qui sont fournis avec un indice de confiance, comme la probabilité de survie. Après avoir démarré avec les cancers du poumon, les cancers couverts intègrent maintenant les leucémies, les mélanomes, ceux du pancréas, des ovaires, du cerveau, du sein et du colon.

Dans cette application, Watson bat l'homme dans la force brute : il compulse notamment des bases de données de recherche en oncologie pour aider les cancérologues. Mais d'où viennent ces données ? Fait-il progresser la recherche ? Indirectement oui car il va alimenter ces bases de données qu'il utilise avec des résultats de traitement saisis par les praticiens.

Par contre, il ne fait pas directement progresser la recherche sur les cancers. Il ne faut pas oublier que les articles scientifiques exploités ont chacun nécessité de 3 à 7 années de recherche réalisées par plusieurs chercheurs ! C'est un travail considérable. Watson utilise les résultats de la recherche existante qui s'appuie sur des expériences in-vitro et in-vivo, que l'on ne sait pas encore simuler numériquement, et les résultats statistiques associés. Bref, on a encore besoin de chercheurs ! Pour automatiser ce processus, il faudrait passer par plusieurs stades d'évolution de l'IA : ajouter la dimension créative et conceptuelle, automatiser des tests in-vitro et in-vivo avec des robots et en dernier lieu, bien plus tard, réaliser ces tests in-silico quand les algorithmes et la puissance de calcul le permettront.

Tout cela est bien merveilleux mais le marketing d'IBM autour du cancer est un peu trompeur et ses prouesses largement survendues.

Les annonces évoquées ci-dessus ne sont pas vraiment éprouvées ni déployées à grande échelle[560]. En 2018, on pouvait même constater une grosse déconvenue sur le sujet même si IBM avait du mal à l'admettre[561].

En août 2018, IBM contre-attaquait en indiquant que Watson fournissait une aide appréciable aux cancérologues[562]. Watson for Oncology serait utilisé dans 230 hopitaux avec 84 000 patients traités sur H1 2018.

---

[559] Dans How to Make More Published Research True, John Ioannidis indiquait en 2014 que 85% des ressources des chercheurs sont gaspillées et leurs résultats publiés sont faux ou exagérés. Dans Raise standards for preclinical cancer research, Glenn Begley et Lee Ellis indiquaient en 2012 que 90% des résultats de 53 études majeures dans le domaine du cancer n'étaient pas reproductibles. Donc, si elles sont utilisées par Watson, il ne peut pas en sortir grand-chose d'utile ! Voire, cela peut même être dangereux. Enfin, dans Believe it or not: how much can we rely on published data on potential drug targets?, Florian Prinz, Thomas Schlange et Khusru Asadullah indiquaient en 2011 que 79% des résultats de 67 travaux de recherche en cancérologie et cardiologie n'étaient pas reproductibles chez Bayer. Qui plus est, les recherches qui donnent lieu à des résultats négatifs sont bien moins publiées celles qui sont concluantes. Ce sont toutes ces études qui alimentent Watson for Oncology ! Le biais statistique qu'elles induisent est énorme ! Source de cette liste : Bio-Modeling Systems - The Mechanisms-Based Medicine Company de Manuel Gea, juillet 2017. D'où l'intérêt d'initiatives comme le centre METRICS de Stanford, qui vise à faire de la méta-recherche, donc d'auditer les pratiques des chercheurs pour les améliorer.

[560] Voir IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close de Casey Ross., septembre 2017 et Why Everyone Is Hating on IBM Watson—Including the People Who Helped Make It de Jennings Brown en octobre 2017.

[561] IBM's Watson reportedly created unsafe cancer treatment plans, juillet 2018, et Report: IBM Watson delivered 'unsafe and inaccurate' cancer recommendations, juin 2018 fait état de documents internes de l'équipe IBM Watson Health qui indiqueraient que les résultats de Watson for Oncology comprendraient des résultats incorrects de propositions de thérapies dans sa mise en œuvre sur des cas fictifs portant sur 13 typs de cancers au Memorial Sloan Kettering Cancer Center. Ce qui n'est pas étonnant quand on sait qu'il exploite des méthodes probabilistes qui ne peuvent garantir 100% de bons résultats. IBM licenciait en mai 2018 des équipes issues de trois acquisitions dans la santé (**Explorys**, **Truven** et jusqu'à 80% des salariés de **Phytel**) acquises en 2015 et 2016. Ils ont surtout des difficultés économiques à faire sortir l'IA des PoC (proofs of concepts). Certaines sont liées aux limitations des solutions, d'autres aux difficultés classiques de déploiements d'innovations. Et, avec ou sans IA, il est fréquent que des acquisitions de startups par de grands acteurs finissent mal. Voir Layoffs at Watson Health Reveal IBM's Problem With AI, juin 2018.

Heureusement, il n'y a pas qu'IBM Watson dans les usages de l'IA en cancérologie. De nombreux chercheurs investiguent des méthodes innovantes pour améliorer les traitements. Mais leurs résultats sont moins époustouflants car ils concernent des cas très précis et n'ont pas des prétentions généralistes comme IBM avec Watson.

C'est le cas de cette nouvelle méthode de traitement des glioblastomes, ces cancers du cerveau très difficiles à traiter, avec un ciblage précis des chimiothérapies via un système à base d'apprentissage par renforcement qui optimise à la fois l'efficacité du traitement et minimise le nombre de doses. Le modèle créé par des chercheurs du **MIT Media Lab** a été entraîné avec des tests cliniques sur 50 patients et 20 000 tests pour alimenter l'apprentissage par renforcement. Le modèle entraîné a été ensuite testé sur 50 autres patients "simulés". Reste à le faire sur de véritables patients[563] !

Des méthodes équivalentes sont développées pour optimiser les radiothérapies. C'est le cas de travaux de l'**Université de Toronto**[564] qui raccourci le temps de préparation des séances. Ils utilisent du machine learning à base de PCA (principal component analysis, pour déterminer des variables clés) et de régressions pour analyser un historique de données de radiothérapies et créer des plans de séances optimisés. Le système a été évalué sur 217 patients atteints de cancers de la tête et du cou. La startup **TheraPanacea** (2017, France) est positionnée sur un créneau voisin.

Enfin, citons **Sophia Genetics** (2011, Suisse, $63M) qui piste les mutations de l'ADN des patients de plus de 400 hôpitaux pour améliorer la compréhension de cancers et de certaines maladies rares et la création de traitements personnalisés. Ils font appel au machine learning dans leurs processus.

Chirurgie

La chirurgie fait évidemment de plus en plus appel à des robots. Nombre d'entre eux répètent des gestes de chirurgiens à distance comme les Da Vinci d'**Intuitive surgical** (1999, USA) qui sont spécialisés dans les opérations de l'abdomen et sont déployés depuis plusieurs années. Seuls certains ont une véritable IA lorsqu'ils sont autonomes et exploitent par exemple le résultat d'imagerie médicale réalisée en temps réel sur les patients.

L'IA peut aussi servir à former les chirurgiens. C'est ce que propose **Insimo** (2013, France) qui développe un simulateur réaliste d'organes à base d'IA et d'outils de simulation.

De manière encore plus avancée, **Cambridge Bio-Augmentation Systems** (2015, UK) utilise du machine learning pour comprendre le fonctionnement des nerfs de patients pour leur greffer des membres bioniques. Leur Prosthetic Interface Device (PID) est adapté à l'ajout d'un bras artificiel.

Enfin, les anesthésistes pourront aussi bientôt exploiter l'IA pour mieux anticiper les complications des patients, notamment les risques d'hypoxémie, en fonction des nombreux paramètres qui les caractérisent. C'est l'objet du projet de recherche Prescience de l'**Université de Washington** (dans l'Etat du même nom, à Seattle)[565].

Diabète

Les solutions aidant les diabétiques à suivre leur traitement et à équilibrer leur dose d'insuline (pour le type 1) et leur alimentation sont très nombreuses.

---

[562] Voir IBM pushes back on negative Watson Health stories, août 2018.

[563] Voir Reinforcement learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regiment Selection de Pratik Shah et Gregory Yauney du MIT Media Lab, 2018 (65 pages).

[564] Voir Knowledge - based automated planning for oropharyngeal cancer, 2017 (67 pages).

[565] Voir Machine-Learning Can Help Anesthesiologists Foresee Complications de Marc O'Reilly, octobre 2018 ainsi qu'une version préliminaire de leur publication dans Anesthesiologist-level forecasting of hypoxemia with only SpO2 data using deep learning, décembre 2017 (6 pages). La méthode associe des Convnets et des réseaux à mémoire de type LSTM.

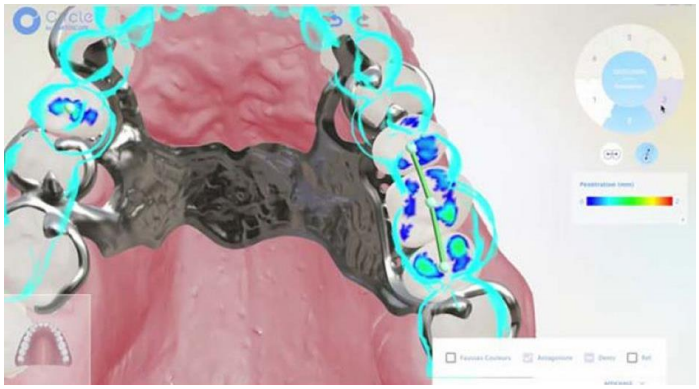
Diabeloop (2015, France, 13,5M€) finalise la mise au point de DBLG1, une solution complète de bout en bout, intégrant un capteur de glycémie G6 en continu d'origine Dexcom, une pompe à insuline d'origine Cellnovo (2002, UK, $102,5M) ou Kaleido (2015, USA, $165M) et un boitier dédié faisant tourner un logiciel de suivi pour ajuster automatiquement le dosage d'insuline, exploitant des briques d'IA. Le DBLG1 a obtenu l'agrément CE en novembre 2018, ouvrant la voie à une commercialisation prochaine. Ce n'est pas la seule solution de gestion du cycle glycémique en boucle mais semble être la seule qui exploite l'IA.

Cognitive Scale (2013, USA, $40M) a créé la solution Cognitive Clouds qui est proposée aux adolescents atteints de diabète type 1 pour les aider à se réguler, en intégrant les aspects médicaux (prise d'insuline, suivi de glycémie), d'activité physique et d'alimentation. Il y a des dizaines de startups qui visent le même marché et avec plus ou moins de bonheur. Très souvent, elles méconnaissent le fonctionnement des diabétiques dans la régulation de leur vie et leur segmentation.

Diabnext (2018, France) est une startup à cheval sur la France, les USA à Boston et Taïwan qui est positionnée sur le même créneau que Cognitive Scale.

Orthodontie

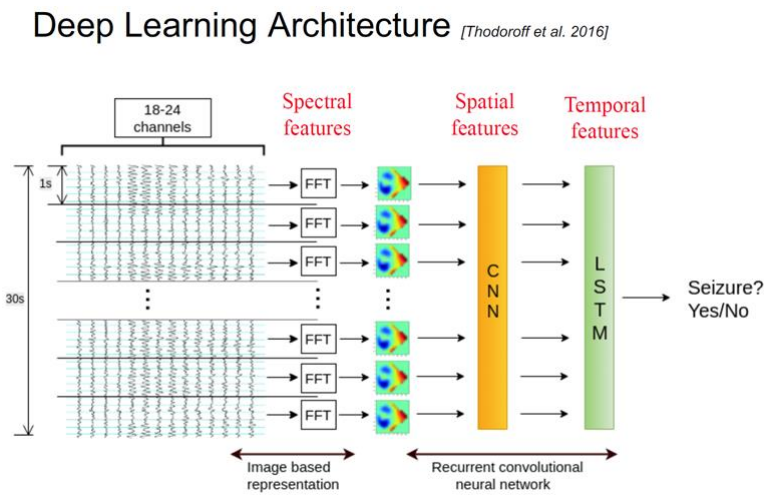
Biotech Dental (1987, France) a lancé en mars 2018 Lucy, un logiciel de simulation biomécanique pour la création de prothèses dentaires. C'est plutôt un logiciel de modélisation et de CAO que d'IA. Mais il propose des traitements qui sont issus d'un entraînement exploitant des études de cas précises. Il s'appuie visiblement sur les travaux d'**Anatoscope** (2015, France) qui fait de la simulation 3D (vidéo). Les prothèses sont ensuite imprimées en 3D (vidéo).



Complémentaire à la solution de Biotech Dental, **Neocis** (2012, USA, $20,9M) a développé le robot de pose d'implants dentaires Yomi. Vous êtes partants pour le tester ? Pour vous rassurer, il est piloté par le dentiste et ne fonctionne pas en roue libre. Reste à savoir si ce robot contient véritablement des briques d'IA !

Neuro et psy

La détection d'épilepsie fait partie des travaux du FAIR, le laboratoire de recherche de **Facebook**, avec de l'apprentissage par renforcement qui associe une analyse spectrale (CNN) et temporelle (LSTM)[566]. Le projet est piloté par Joëlle Pineau, qui dirige la branche de Montréal de ce laboratoire. Le principe consiste à générer une neurostimulation artificielle qui évite le déclenchement inopiné d'épilepsie chez les patients.



Textpert (2015, USA) a développé son « Empathetic Virtual Interface » avec des chercheurs d'UCLA et de l'USC. Il s'agit d'un chatbot de diagnostic psychothérapeute. Il devrait être au point d'ici fin 2019. Mais l'histoire ne dit pas s'il passera le test de Turing à cette échéance. Il ne coutera que $10 par mois, soit beaucoup moins qu'un psychothérapeute, en tout cas aux USA.

Dans un projet de recherche européen Horizon 2020 anglo-japonais, le robot **Nao** peut former les enfants autistes à bien reconnaître les émotions (vidéo)[567].

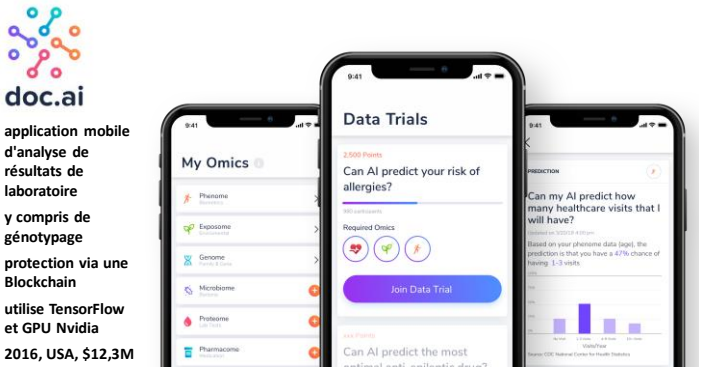
Systèmes de santé

Les solutions de systèmes de santé couvrent des besoins divers : le suivi de l'observance des traitements, l'évitement d'erreurs de prises de médicaments, les robots pour s'occuper des personnes âgées, le suivi des dépenses de santé ou l'optimisation des ressources des hôpitaux et praticiens. Ils génèrent de gros volume de données, d'où les nombreux cas d'usage potentiels de l'IA.

En voici quelques exemples, toujours pris dans l'univers florissant des startups.

Télémédecine

Doc.ai (2016, USA, $12,3M) propose une application mobile et webapp comprenant un médecin virtuel dédié à l'analyse de résultats de laboratoires, y compris de génotypage, ce qui n'empêche pas, ensuite, d'aller voir un médecin et surtout un spécialiste (vidéo). Ils utilisent aussi une Blockchain pour gérer la traçabilité des examens biologiques et leur partage dans des communautés de patients pour des tests cliniques ou autres besoins.

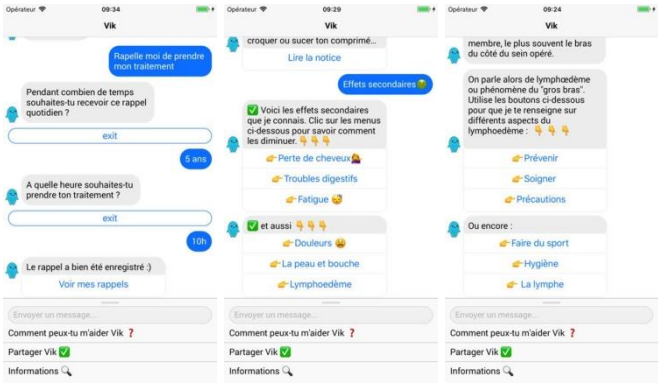


[566] Vu dans Adaptive treatment of epilepsy via reinforcement learning de Joëlle Pineau, 2017 (35 slides).

[567] Vu dans Humanoid Robot Teaches Autistic Kids to Recognize Emotions, juillet 2018.

Wefight (2017, France) a créé Vik Sein, un chatbot de suivi du cancer du sein pour les patientes (vidéo).

Il a été codéveloppé avec le réseau de proximité « Mon réseau cancer du sein ». Ce chatbot semble est assez riche fonctionnellement avec un suivi des traitements, des conseils et plusieurs types d'interactions avec les patientes.



Corti (2016, Danemark) a développé un outil d'analyse de la voix des patients ou de leur entourage qui appellent les services d'urgence pour détecter notamment les pathologies cardiaques et mieux gérer les urgences. C'est original et bien utile ! L'outil Orb alimente l'écran du dispatcher des urgences en l'alimentant avec les informations détectées dans la voix, y compris issues du speech-to-text. Orb utilise des réseaux convolutionnels (vidéo).

Ophtalmologie

Panda Guide (2015, France) propose un système dédié aux aveugles qui se positionne autour du cou et est complété d'oreillettes audio. La partie IA tourne dans le smartphone en mode offline avec un modèle de vision entraîné sur serveur, capable de rcconnaître un millier d'objets de la vie courante.

OrCam Technologies (2010, Israël, $86,4M) apporte la vision aux mal voyant via une caméra reliée à un système de reconnaissance d'objets qui décrit les scènes de manière vocale. C'est une startup israélienne. On a ici un mélange de computer vision et de text to speech. La caméra miniature se positionne sur des lunettes traditionnelles et comprend un écouteur.

Autres

Aux USA, le projet **Deep Patient** peut prédir l'apparition de 78 pathologies via un entraînement exploitant les dossiers médicaux de 1,6 millions de patients de l'hôpital MtSinai de New York étalés sur plusieurs décennies[568]. Il a entre autres bénéficié de la loi *Health Information Technology for Economic and Clinical Health Act* votée en 2009, juste avant l'Affordable Care Act (Obamacare). Elle encourageait les hôpitaux et médecins à adopter des systèmes de dossiers médicaux interopérables. 84% des hôpitaux avaient emboité le pas en 2017.

MedAware (2012, Israël, $10,4M) fournit une solution qui permet d'éviter les erreurs de prescription médicamenteuse en temps réel pour les médecins. Avec des morceaux de big data et de machine learning dedans qui exploitent notamment des bases de données médicales d'historiques de patients.

[568] Voir Deep Patient: Predict the Medical Future of Patients with Artificial Intelligence and EHRs de Riccardo Miotto, 2017 (51 slides) ainsi que Dans Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis, février 2018 (16 pages) qui fait le point des méthodes à base d'IA d'exploitation des dossiers médicaux aux USA.

Hindsait (2013, USA) a une solution en cloud servant à identifier les déviations dans les dépenses de santé. Cela sert donc surtout aux financeurs des systèmes de santé que sont les assurances publiques, privées et les mutuelles. Ca fait moins rêver le patient ! Ce genre de solution fait du prédictif multicritère à base de machine learning.

Predical (2015, France) fait partie de ces startups de la silver economie et des objets connectés qui veut faire du prédictif avec de l'apprentissage d'IA par la détection des routines de la vie et les anomalies associées.

Senscio Systems (2009, USA, $7,9M) commercialise Ibis Health Management Solution qui est censé réduire les coûts de santé pour des organismes payeurs en améliorant le suivi de maladies chroniques. Le principe ? Détecter les déviations des patients par rapport aux prescriptions des médecins. Cela a l'air de supposer que le patient utilise un grand nombre d'objets connectés pour suivre sa tension, sa glycémie, son activité physique et son poids, entre autres paramètres vitaux. **Innovaccer** (2014, USA, $43M) semble être sur le même marché.

Somatix (Israël) a développé un système de détection des gestes via des wearables. Leurs basés aux USA sont des hôpitaux. La solution sur le marché pour le moment est SmokeBeat, et permet d'assurer une télésurveillance des patients qui suivent un traitement anti-tabac. Ils développent aussi une solution de suivi des soins des seniors.

macro-eyes (2014, USA) propose un outi original qui permet de mieux gérer la prise de rendez-vous avec les médecins tout en réduisant les risques de no-show des patients.

En France, Emmanuel Bacry (CNRS) exploite les données des parcours de santé de la CNAM pour d'identifier des effets secondaires indésirables de médicaments comme pour l'antidiabétique oral Pioglitazone qui a été retiré du marché en 2011 à cause de son incidence sur le cancer de la vessie[569]. Les données utilisées proviennent du SNIIRAM (Système National d'Information Interrégimes de l'Assurance-maladie) qui comprend un milliard de feuilles de soins de 65 millions d'assurés.

Finance

Le secteur de la finance est un terrain très favorable à l'usage de techniques de machine learning, dans tous ses métiers[570], du front office au back office en passant par la relation clients via des chatbots, l'analyse de risques et l'optimisation de portefeuille, sans compter les techniques de base utilisées depuis longtemps comme la reconnaissance automatique de l'écriture manuscrite dans les chèques. L'objectif est toujours d'optimiser les opérations, d'en réduire les coûts, de personnaliser les offres et d'améliorer la relation client.

Les services financiers donnent lieu à la création d'un grand nombre de startups qui fournissent quelques indications des usages prometteurs de l'IA dans les métiers de la finance (voir la cartographie de CBInsights plus loin, qui date de 2017 et n'a pas été actualisée depuis).

Le marché bancaire est très verticalisé et a plutôt bien résisté aux coups de boutoir des startups depuis plus de 20 ans, malgré une insatisfaction chronique de certains segments de clients. La banque directe se développe lentement, surtout en France.

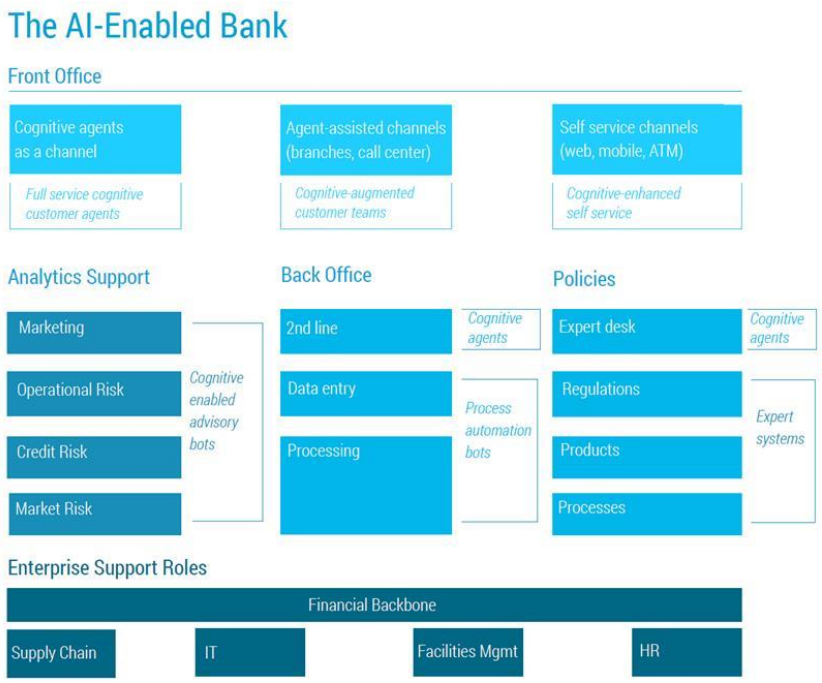
Avec ou sans IA, les Fintechs ambitionnent de disrupter le marché à tous les étages (mobilité, Bitcoins, crowdfunding, etc) mais sont encore très focalisées sur les moyens de paiement et moins sur la gestion de comptes et la commercialisation de produits financiers.

---

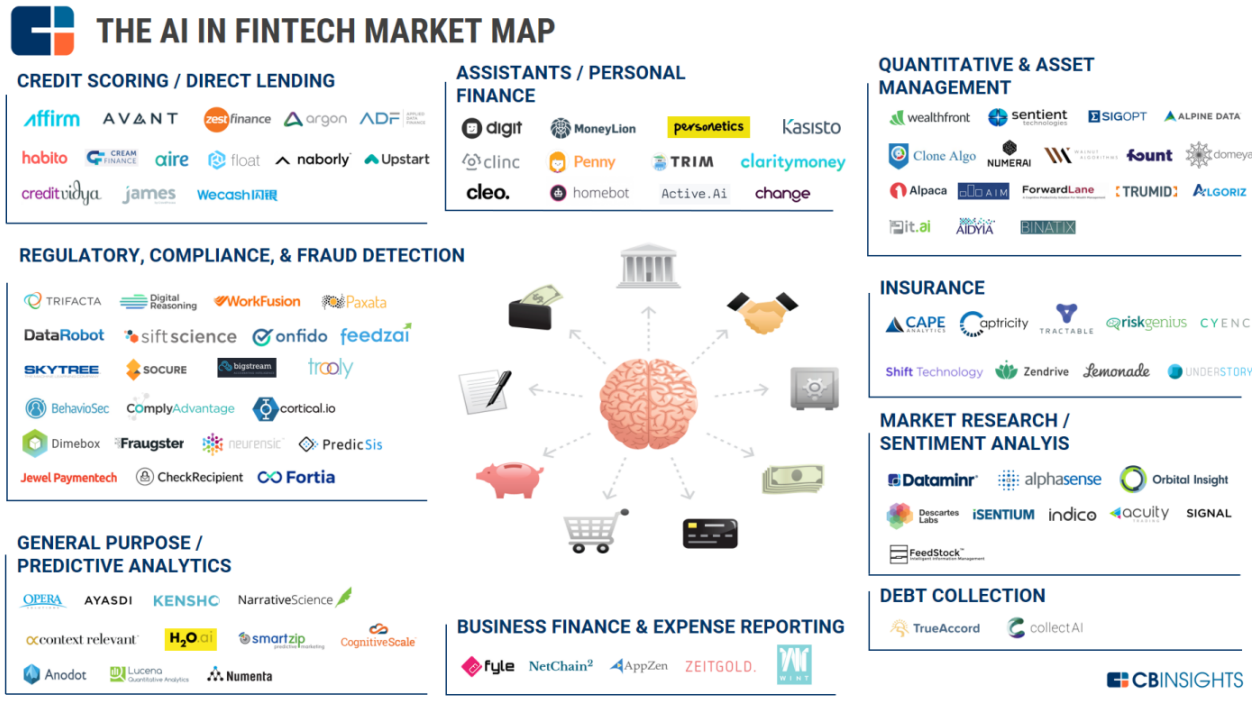
[569] Voir Big data : premier succès dans l'alerte sanitaire, janvier 2018.

[570] Le schéma vient de Tomorrow's AI-Enabled Banking de IPSoft.

La banque de détail a d'ailleurs déjà traversé sa révolution numérique en transférant une part du travail salarié des agences et du back-office vers les clients se débrouillant par eux-mêmes sur les guichets automatiques et autres applications web et mobiles. Cela les a amenés à revoir le métier des agences, plus positionnées sur la commercialisation de produits financiers. Cette transformation déjà réalisée correspond assez bien à ce qui se produira dans les années qui viennent dans les cabinets d'expertise comptable, voir chez les notaires et certains avocats.



L'IA pourrait-elle accélérer la mutation du secteur bancaire ? Est-elle un facteur qui peut faire évoluer la notion de confiance, critique dans le secteur, d'un type d'acteur à un autre [571]? Pour l'instant, rien ne semble l'indiquer malgré les coups de boutoir des startups et des cryptomonnaies.



[571] Le problème étant que le grand public n'a plus confiance en grand-chose. Tout le monde en prend pour son grade : les politiques, les médias et les entreprises. Voir le baromètre mondial de la confiance 2017 d'Edelman.

Marketing

Le ciblage de clients est proposé par **DataFox** (2013, USA, $11,8M) qui utilise force big data et machine learning couplé au CRM et à la gestion de clients entrants pour cibler et traiter les bons clients avec les bonnes offres.

Optimisation d'investissements

L'exploitation de l'IA couvre toutes les solutions d'optimisation de gestion des investissements, surtout boursiers. La majorité des solutions du marché côté startups sont b2c et quelques-unes sont b2b. Les solutions b2c ciblent en premier lieu les audiences adultes les plus jeunes, ces fameux millenials.

L'optimisation d'investissements boursiers est proposée par **AdvisorEngine** (2014, USA, $54,3M) avec son logiciel en cloud de gestion d'investissements personnels pour les conseillers en gestion de patrimoine et leurs clients (*ci-contre*). Ils utilisent du machine learning mais ne précisent pas comment dans leur communication comme nombre de startups de ce secteur, ni quelles données servent à l'entraînement des IA.

WealthArc (2015, USA, $1,5M) propose une solution voisine également destinée aux conseils en gestion de patrimoine (vidéos de démonstration).

Acorns (2012, USA, $152M) automatise la création de portefeuilles financiers selon le niveau de risque souhaité. Ils ont associé leur offre à un système de cashback avec quelques retailers et marques cmme Apple, Airbnb, DirecTV. Cela permet indirectement de faire des économies à placer à l'occasion des dépenses associées.

Stash (2015, USA, $116,3M) permet de gérer ses investissements personnels de manière sélective en privilégiant par exemple les sociétés qui protègent l'environnement. La startup gère aussi vos comptes en banque et votre carte de crédit.

La gestion financière en ligne est proposée par **Wealthfront** (2008, USA, $204M) et **Betterment** (2008, USA, $275M), deux startups extrêmement bien financées et qui ne ciblent visiblement que le marché US. La première a une offre orientée « scénarios » qui permet de planifier le financement des événements de la vie (achat logement, voyages, études des enfants, retraite, …).

wealthfront

gestion financière en ligne
exploitation de machine learning pour optimiser les investissements par classe
startup US créée en 2011
$129m levés

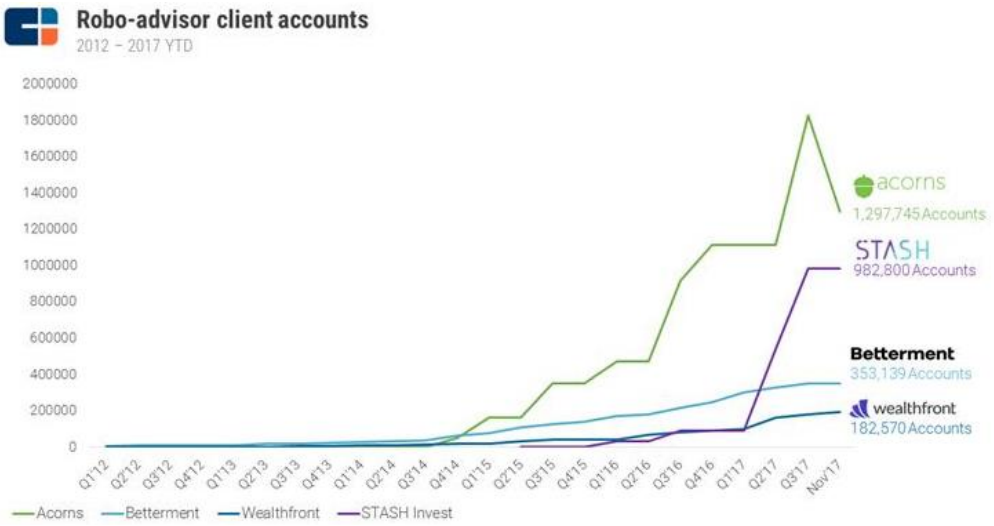
Pefin (2015, USA) gère également aussi la finance personnelle des foyers de tous niveaux de patrimoines, aussi bien pour planfier les investissements que pour gérer ses comptes au plus juste, selon les situations. La startup indique surveiller 2 millions de données pour optimiser ses recommandations : dépenses personnelles, indicateurs économiques divers, fiscalité, couverture santé, etc (vidéo).

gestion financière en ligne pour les familles, entièrement gérée par logiciel et l'IA

IA à base de réseaux de neurones

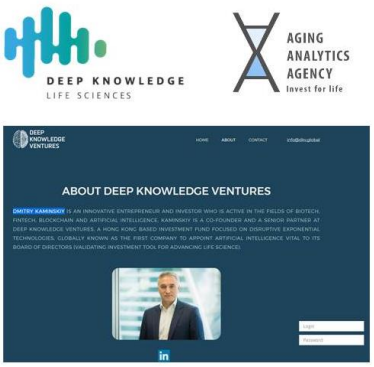
startup basée à New York

CBInsights montre comment ces acteurs que nous venons de citer se positionnent dans la course à la captation de clients, avec Acorns et Stash en tête[572]. Ces startups ont levé des montants très élevés : $747M en tout !



On trouve même des solutions pour identifier des startups dans lesquelles investir. Le fonds d'investissement de Hong-Kong **Deep Knowledge Venture** spécialisé dans la santé communiquait avec fracas en 2014 sur le fait que l'un des membres de son comité d'investissement était une IA. C'est en fait une solution extrêmement pointue, développé par Aging Analytics (UK), spécialisé dans la recherche sur les technologies de lutte contre le vieillissement.

Mais indiquer qu'une IA a une place dans un board est aussi stupide que si un cabinet d'expert comptable indiquait qu'un PC équipé d'un tableur Excel était un de ses employés. Aussi sophistiquée soit-elle, une solution d'IA reste un outil d'aide à la décision, comme dans n'importe quel autre processus de décision qui s'appuie sur la rationnalité de données. Mais il faut toujours y ajouter un peu d'intuition et de connaissance humaine !



fonds d'investissement de Hong Kong créé par le moldave Dmitry Kaminskiy

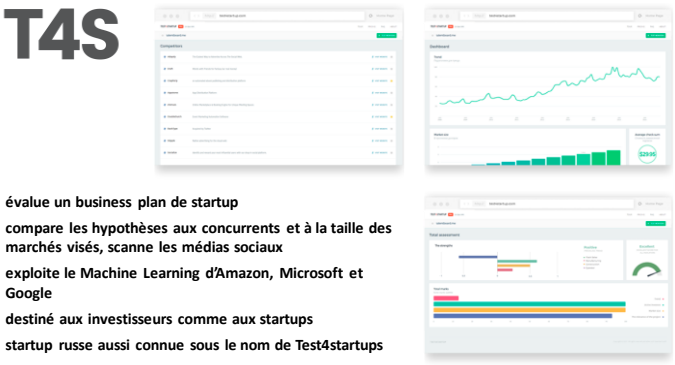
IA Vital (Validating Investment Tool for Advancing Life Sciences) qui siège dans le comité d'investissements

aide à la décision d'investissements dans les startups de technologies exponentielles de la santé

développé par Aging Analytics (UK), spécialisé dans la recherche sur les technologies de lutte contre le vieillissement

[572] Ici : Acorns Teardown: The Most Popular Robo-Advisor Faces A Fierce Fight As It Goes 'Upmarket', novembre 2017.

Le russe **TalentBoard** et sa solution Test4startups (ou T4S) aide de son côté les startups à valider leur business plan et leur marché. Le système analyse le marché visé, les concurrents, la liste des investisseurs potentiels et produit une évaluation du projet. Il faut prendre cela avec quelques grains de sel car cela ne suffit évidemment pas à prendre une décision éclairée d'investissement dans une startup, surtout d'innovation « de rupture ».



Le fonds suédois **EQT Venture** gère un fonds d'investissement dans les startups de 566M€. Il fait partie de la société d'investissement EQT qui gère $38B d'actifs.

EQT Ventures exploite sa propre solution d'IA dénommée Motherbrain pour faire ses choix d'investissements et surtout, détecter les startups les plus prometteuses qui sont « sous le radar » et, ce qui est le plus important, avant les autres fonds[573]. L'IA exploite une myriade d'informations comme les financements de startups, les classements sur Internet et l'activité dans les réseaux sociaux plus des données générées par ses propres analystes. Le machine learning exploité est non supervisé pour identifier des tendances et supervisé pour exploiter des données déjà labellisées. Motherbrain est aussi exploité par le fonds pour accompagner les startups dans lesquelles il investi[574].

Mattermark (2012, USA, $17,2M) est un fournisseur de base de données d'entreprises servant à la détection de prospects pour la vente en tout genre et pas seulement pour les investissements par les sociétés de capital risque. Là encore, il faut regarder de près les données qui alimentent leur système et prendre l'ensemble avec des pincettes.

Il faut aussi compter avec **Kensho Technologies Inc** (USA) qui s'appuie sur l'IA en général pour fournir des outils d'analyse aux investisseurs dans les banques.

Et ne j'ai pas cité les projets et startups qui cherchent à prédire les cours de la bourse ou d'indices boursiers à base d'IA, comme **QuantCube** (France, $5,1M)[575] ! Nous avons aussi **Cognivi Labs** (2016, USA, $3,1M) squi utilise force machine learning à base de calcul de la confiance des consommateurs pour prédire les évolutions des cours de bourse de grandes entreprises, avec une étude de cas sur Facebook et une autre sur Shechers. A prendre avec des pincettes car ils ne publient évidemment pas les cas où leur prévision n'a pas bien fonctionné ! Les biais statistiques, toujours les biais statistiques !

[573] Je suis évidemment très réservé sur ce genre de méthode pour trouver les véritables pépites. Les données génèrent ce que l'on appelle le biais du rétroviseur : elles ne permettent pas de voir en avant ! Qui plus est, les facteurs clés de succès d'une startup sont des signaux faibles difficiles à détecter. Les données permettent cependant sans doute d'éliminer les plus mauvais projets, même si un coup d'œil sur le dossier et sur l'équipe permet aussi de le faire. C'est la détection des meilleurs projets qui est difficile, surtout lorsque l'on sait que le succès des meilleurs dépend aussi de la chance et de phénomènes sociétaux. Même si ces derniers peuvent être détectés dans les réseaux sociaux.

[574] Les fonds américains **SignalFire** et anglais **InReach Ventures** utilisent aussi du machine learning pour choisir les startups dans lesquelles ils investissent.

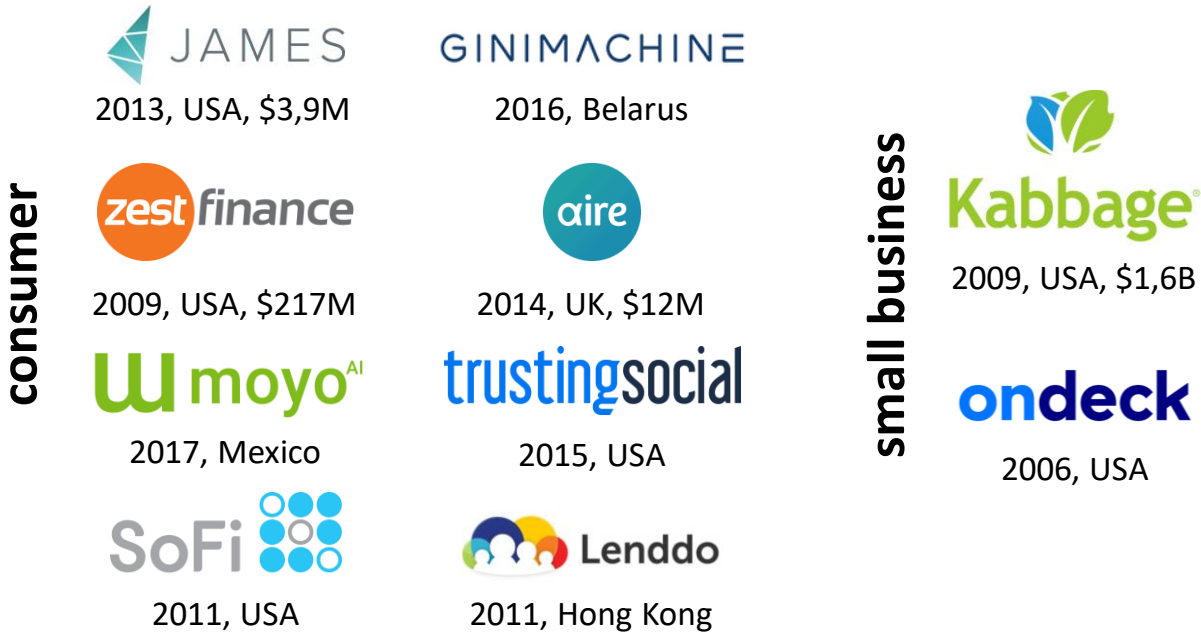
[575] Voir aussi Forecasting the Stock Market Index Using Artificial Intelligence Techniques de Lufuno Ronald Marwala (166 pages).

Gestion des risques

La gestion des risques porte sur le credit rating[576] d'emprunteurs basé sur les informations disponibles sur les réseaux sociaux avec **TrustingSocial** (2015, USA), **Lenddo** (2011, Hong Kong, $14M) ou **Kreditech** (2012, Allemagne, $497M), ce qui n'est pas sans poser diverses questions sur le respect de la vie privée.

De nombreux services de crowlending tels que **Kabbage** (2009, USA, $1,6B) qui cible le marché des PME aux USA et garantit des prêts qui font en moyenne $200K, **OnDeck Capital** (2006,USA) qui cible aussi les PME, **SoFi** (2011, USA) qui gère des prêts pour étudiants, **Lending Club** (2007, USA) et **ZestFinance** (2009, USA, $217M) font aussi appel au machine learning pour le credit rating, ce dernier en se focalisant sur les emprunteurs qui ont une faible empreinte en ligne. **OakNorth** (2015, UK, $448,5M) fait des prêts aux PME dont le risque est aussi évalué avec de l'IA avec leur solution ACORN.

Le *credit rating* à base d'IA est aussi proposé par **James** (2013, USA, $3,9M), **Aire** (2014, UK, $12M) qui se focalise sur les jeunes emprunteurs sans 'casier' financier. **GiniMachine** (2016, Belarus) a une offre voisine. **Upstart** (USA, $85,7M) propose des prêts aux particuliers. Le machine learning permet de générer les taux d'intérêts et autres conditions de prêts automatiquement.



Smart Finance (2013, Chine) demande aux utilisateurs d'accéder à leurs données mobiles. Le machine learning vient à la rescousse pour identifier des corrélations entre des centaines de paramètres de la vie numérique mobile des mobinautes et leur fiabilité d'emprunteur. Cela va jusqu'à récupérer le niveau de charge de la batterie. Une faible charge chronique serait un paramètre corrélé avec une propension à ne pas rembourser ses prêts ! Voilà de quoi créer de beaux biais des données !

LendUp (2011, USA, $361M) propose des prêts aux particuliers en optimisant leurs taux d'intérêts de prêts. C'est une forme de système de micro-sub-primes, attribuant des crédits à des particuliers qui ne peuvent pas accéder aux crédits dans les circuits traditionnels. Comme c'est une activité plus risquée que les crédits traditionnels, la startup doit se couvrir avec des algorithmes qui limitent les risques en question !

---

[576] Voir Application of Artificial Intelligence Techniques for Credit Risk Evaluation de Ahmad Ghodselahi et Ashkan Amirmadhi qui décrit une méthode de credit rating à base d'arbres de décision, de machine learning et deep learning exploitant une dizaine d'agents différents. Voir aussi Credit Scoring Models by AI Companies: A Comprehensive Guide, Credit Scoring Using Machine Learning, 2013 (381 pages) et Artificial intelligence and machine learning in financial services, 2017 (45 pages).

L'IA détecte les clients potentiels qui ont des comportements « sains » comme…le fait de bien rembourser ses crédits et d'avoir un budget équilibré ! On peut se demander pourquoi il faut de l'IA pour déterminer cela !

En mode b2b, il est aussi réalisé par **JPMorgan Chase** depuis 2017 avec sa solution COiN, qui utilise le machine learning pour analyser leurs 12 000 contrats de crédit commercial annual en quelques secondes, faisant sur le papier gagner 360 000 heures par an à la banque d'investissement.

RCI DIAC, filiale du groupe **Renault** s'appuie sur **Score4Biz** (2012, France) pour déterminer en temps réel un score d'octroi de crédit. Elle a déployé une solution de machine learning basée sur des travaux de recherche issus de Telecom ParisTech et de l'ENS Paris Saclay. Les engagements de services comprennent un temps de réponse de 250 ms en moyenne avec un engagement contractuel de moins d'une seconde.

Enfin, l'évaluation du risque doit être aussi faite côté client en b2b. C'est la mission de **Neuroprofiler** (2016, France) qui propose un jeu en ligne qui permet de qualifier le profil d'investisseur financier de clients, pour respecter la règlementation financière européenne MiFIDII. Le tout s'appuie sur du deep learning qui associe les réponses au jeu à un profil type financier en terme de capacité et de compréhension de prise de risque.

Détection de fraudes

La détection de fraude est un cas d'application classique du machine learning. Les fraudes sont détectées en collectant un maximum d'information sur les payeurs et en identifiant les « patterns » de mauvais payeurs. C'est ce que propose **Sift Science** (2011, USA, $106,6M) avec une offre de sécurisation généraliste destinée aux banques et commercants et aussi **Riskified** (2012, Israël, $63,7M), qui est focalisé sur les sites marchands.

DataVisor (2013, USA, $54,5M) fait de la détection d'anomalies dans les transactions financières à base de machine learning et **Quantexa** (2016, UK, $23,3M) aide à lutter contre la criminalité financière, le blanchiment d'argent sale et à réduire les risques de crédit. Shell Oil fait partie de leurs premiers clients.

La banque **Santander** utilise la reconnaissance vocale pour sécuriser les transactions, avec la solution de la startup **Fonetic** (2006, Espagne) qui analyse les émotions dans les conversations téléphoniques.

Les banques doivent aussi passer au peigne fin toutes les transactions de plus de $10K pour détecter le blanchiment d'argent sale. Là encore, il faut faire appel à du machine learning voire du deep learning pour trier les centaines de milliers de transactions. C'est ce que propose de faire la startup **Simularity** (2011, USA) qui aide à détecter les anomalies dans de nombreux marchés verticaux, dont la finance[577].

Deux startups spécialisées dans les applications de gestion de la conformité des transactions, l'Américaine **Lucid** (2004, USA, $15M, acquise par Acuity Brands en février 2018) et **Feedzai** (2009, USA/Portugal, $76M) utilisent toutes deux le machine learning pour détecter 80% des fraudes.

Enfin, citons **Coinbase** (2010, USA) qui gère un portefeuille électronique de Bitcoin et s'appuie sur de l'IA pour éviter la fraude et l'usurpation d'identité[578].

---

[577] Voir leur intéressant livre blanc Artificial Intelligence (AI) for Financial Services.

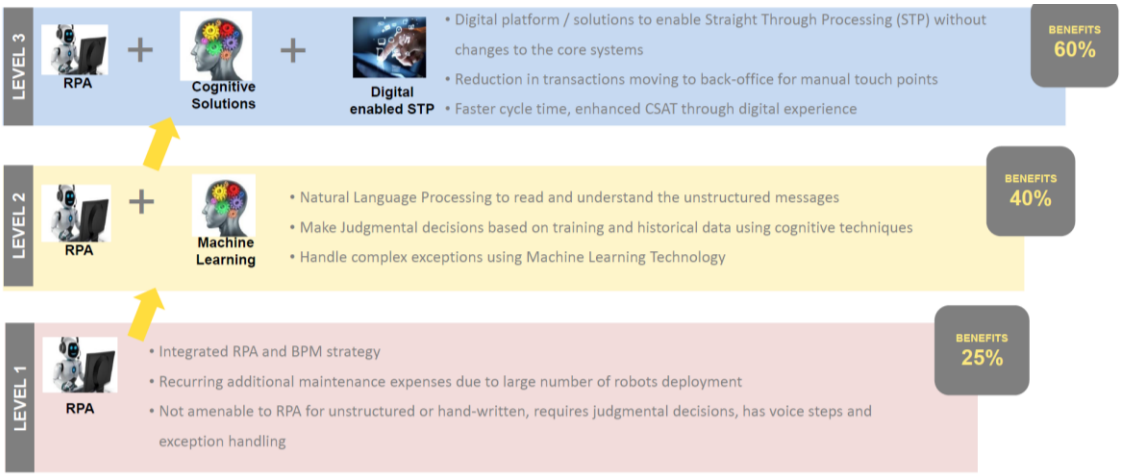
[578] Voir AI at scale at Coinbase, septembre 2018.

Robotic Process Automation

Depuis 2015, la **Robotic Process Automation** décrit les outils d'automatisation des processus internes des entreprises couvrant la finance et le marketing. Elle consiste à permettre à des agents à base d'AI de naviguer par eux-mêmes dans les différentes applications de l'entreprise afin de mener des tâches prédéfinies comme la collecte de documents.

L'IA permet en théorie à ces agents d'évoluer par eux-mêmes pour ingérer de nouvelles règles. Les études de cas se font jour depuis 2015[579].

L'un des impacts de cette robotisation des processus sera de réduire l'emploi dans les entreprises concernées et surtout chez leurs sous-traitants, et notamment en Inde pour les entreprises anglo-saxonnes[580].



UiPath (2005, USA, $408M) est positionné exactement sur ce créneau-là[581]. La solution se découpe en trois parties : UiPath Studio qui permet de décrire les processus business de l'entreprise, et aussi d'enregistrer des sessions d'accès à des applications, UiPath Robot qui gère l'automatisation des processus et UiPath Orchestration qui permet de gérer le robot et l'orchestration (vidéo). Il va sans dire que cette automatisation des processus n'a rien de magique et qu'elle requiert beaucoup de paramétrage manuel. C'est actuellement le leader mondial de la RPA. Cette forme de RPA reprend les anciens concepts des solutions de gestion de workflow.

Derrière UIPath se battent en duel un bon nombre de sociétés difficiles à départager telles que **Automation Anywhere** (2003, USA, $250M), **Kryon Systems** (2008, USA, $13M), **Kofax** (1991, USA, qui fait partie de Lexmark depuis 2015), **Blue Prism** (2001, UK, $59M) et **Contextor** (2000, France, 600K€). La RPA sophistiquée doit s'appuyer sur l'exploitation de données non structurées et notamment le langage. D'où l'intérêt de **Recital** (2017, France) qui propose une solution de recherche de texte qui cible la finance et l'assurance. Elle automatise le traitement des mails, extrait des données de documents textuels. Elle a Natixis comme premier client.

---

[579] Voir Robotics Process Automation 6 questions to master it, janvier 2017 de la société de conseil parisienne Ailancy.

[580] Voir Introduction to Robotic Process Automation, a primer, de l'IRPA, Institute for Robotic Process Automation, 2015. Source du schéma : RPA and Beyond de TATA, juin 2017. Voir aussi The robots are ready. Are you? (28 pages), une étude de Deloitte sur l'impact de la RPA dans les entreprises.

[581] Voir Les "robots logiciels" de cette startup roumaine prennent le travail peu qualifié d'employés de bureau, août 2017.

Chatbots

Comme dans le commerce en ligne, les chatbots sont très à la mode et donnent lieu à une pléthore d'offres de startups et de banques en ligne.

Le marché des chatbots financiers se structure en gros en trois types d'acteurs : les startups proposant leur chatbot grand public qui s'interface plus ou moins avec les services bancaires de l'utilisateur, les startups qui proposent des chatbots en marque blanche pour les banques et les banques qui proposent un chatbot à leurs propre clients.

Les startups de la première catégorie sont très nombreuses avec **Cleo** (2015, UK, $13,3M), **Chip** (2016, UK), **Digit** (2013, USA, $36M), **Trim** (2015, USA, $2,2M), **Penny** (2015, USA, $1,2M, acquis par Credit Karma en 2018) et **Dyme.co** (2014, USA, $250K). Il va sans dire qu'il y aura de la casse dans ce secteur, comme dans n'importe quel marché dans lequel s'engouffrent des dizaines de startups faiblement différentiées, et souvent, pas très bien financées. Il est probable que celles qui s'en sortiront le mieux seront celles dont l'écosystème sera le plus dense, avec les interfaces avec les banques, d'autres services financiers voire commercants.

Côté marque blanche, nous avons avec notamment l'Américain **Kasisto** (2013, USA, $28,3M)[582] et son chatbot qui est notamment utilisé par MasterCard, **Finn.ai** (2014, Canada, $3M) et **Personetics Technologies** (2010, Israël/USA, $18M) qui couple un agent conversationnel avec du prédictif sur les besoins des clients.

Et puis, nous avons bien évidemment **IBM Watson** qui peut être mis en musique pour créer des chatbots, avec l'accompagnement des équipes services d'IBM. Il est notamment mis en œuvre en France par **Crédit Mutuel CIC**, non sans quelques réactions négatives des syndicats de salariés inquiet pour l'impact sur l'emploi dans les agences. Le chatbot mis en œuvre dans cette banque est un outil pour 20 000 salariés[583] qui trie 350 000 mails entrants par jour et divers chatbots permettant d'interroger les bases de connaissance de la banque sur ses différents métiers (épargne, assurance auto, assurance santé et prévoyance). L'ensemble est développé par l'équipe de la Cognitive Factory avec 75 personnes dont une moitié provient d'IBM. Le projet aurait couté 40M€ étalés sur 5 ans.

Aux USA, des chatbots ont été lancés par **Bank of America** avec Erica qui est plutôt structuré comme un système de recommandation et est aussi commandable par la voix (*ci-dessous à droite*), et **American Express** sur Facebook Messenger depuis fin 2016 et **Wells Fargo** depuis mi 2017.

Au Royaume Uni, **Barclays** a lancé son Launchpad qui permet d'exécuter des tâches de son application mobile classique en mode dialogue, sans que cela soit d'ailleurs plus efficace. Le chatbot de la **Royal Bank of Scotland** est développé avec IBM Watson. On trouve aussi des chatbots chez **Santander** en Espagne[584] et **Swedbank** en Suède avec son agent conversationnel textuel et vocal développé par l'Américain Nuance.

En Chine, les principales banques comme la **Bank of China** ont un chabot intégré dans **Wechat**. Enfin, la banque **OCBC** de Singapour a créé un chatbot dénommé Emma (tous les prénoms féminins y passent…) spécialisé dans l'attribution de prêts pour la rénovation de logements[585].

---

[582] Avec plus ou moins de bonheur, voir : https://www.wired.com/2016/06/new-banking-ai-now-chatbots/.

[583] Confirmé par une décision de la Cour de Cassation du 12 avril. Voir Cour de cassation, civile, ch. sociale, arrêt du 12 avril 2018 qui a considéré que l'impact du déploiement d'IBM Watson sur les conditions de travail des salariés étaient mineures. Voir aussi cette intéressante étude de l'impact de l'IA sur l'emploi dans la finance Impact of Artificial Intelligence (AI) on Financial Job Market – An Introduction 2018 (44 pages). Le chatbot permettrait de faire gagner 10 minutes par jour aux collaborateurs de la banque de détail.

[584] Voici la source de nombre de ces différents exemples : « Artificial Intelligence in Digital Bankin » de MAPA, novembre 2016.

[585] Source : OCBC bank launches first artificial intelligence powerd home & renovation loan specialist, avril 2017.

En France, nous avons aussi **Orange** qui a lancé en 2017 son offre bancaire avec Orange Bank, qui s'appuie fortement sur une application mobile et l'usage intensif d'IA, notamment dans un chatbot. Orange Bank a été développé, entre autres, autour d'IBM Watson. Le lancement en novembre 2017 a été chaotique et la montée en charge commerciale du service est sujette à caution.

Mais au juste, est-ce que les chatbots fonctionnent bien et sont appréciés des utilisateurs ? Rien n'est moins sûr ! La qualité d'un chatbot dépend surtout des processus qui ont été intégrés dans sa base de connaissances. Ils sont souvent très limités et les capacités de dialogue des chatbot ne vont pas très loin. Un bon chatbot doit laisser la main à un véritable interlocuteur lorsqu'il détecte que la communication ne se déroule pas convenablement et les banques n'ont pas encore mis en place de genre de solution.

Le message marketing mis en avant est toujours ambigu : les chatbot permettent d'améliorer la satisfaction client. Ils servent surtout à réduire les coûts de la banque de détail[586].

Assurance

Comme la finance, le métier d'assureur tourne beaucoup autour de la donnée. L'optimisation de la gestion du risque est encore plus critique que dans les services financiers puisqu'elle fait partie du cœur de métier.

distribution	gestion des risques	gestion des sinistres
optimisation des **canaux de distribution** (ML)	**évaluation personnalisée** des risques et des offres (ML)	**qualification des sinistres** via analyse de photos et des textes de description (DL)
optimisation de l'approche de **nouveaux marchés** (ML)	**ciblage des clients** à risques faibles (ML)	**détection des fraudes** (ML)
recommandations de **packages d'assurances optimisés** (ML)	**stratégies de réduction** ciblée des risques (modes de vie pour la santé, style de conduite pour l'automobile, protection de l'habitat pour la sécurité) (ML, DL)	comparaison et **optimisation des devis** de fournisseurs agréés (ML)
création d'offres de **cross-selling** (ML)		**relation client** accélérée via des chatbots (NLP, DL)
		aide à la **saisie de déclaration** de sinistres (ML)
		prévisions des vagues de sinistres pour optimisation d'allocation du capital (ML)

Dans l'assurance, l'IA intervient dans tout le cycle produit : pour segmenter ses clients, créer des produits, cibler ceux qui présentent le moins de risques, proposer les bonnes offres aux clients, gérer des actions préventives de réduction des risques chez les clients, gérer les expertises et détecter les fraudes[587]. La relation client fait, comme dans les banques, aussi appel aux chatbots[588].

De nombreuses startups se sont évidemment aussi lancées dans ce secteur et en exploitant des briques d'IA.

Distribution

Commençons par la distribution de contrats d'assurance.

---

[586] Voir Bots Aren't Ready To Be Bankers de Forrester, août 2016.

[587] Voir Emerging Technologies Transforming the $4tn Insurance Industry, de CommerzVentures, 2016. Insurance 2030—The impact of AI on the future of insurance de Ramnath Balasubramanian, Ari Libarikian et Doug McElhaney, avril 2018 et AI in insurance Pega (13 pages).

[588] Voir Why Chatbots Are Taking over the Insurance Industry, août 2017.

Zelros (2016, France) propose un chatbot spécialisé dans l'assurance pour la relation client en avant-vente et pour la vente de contrats et la gestion des déclarations de sinistres. Ils auraient comme premiers clients CNP Assurances, MAIF, Natixis Assurance et Axa. Le chatbot est destiné aux collaborateurs des assurances pour leur permettre de servir plus efficacement leurs clients.

Minalea (2015, France) a développé un chatbot de vente de contrats d'assurance, l'Assistant Commercial Intelligent pour les conseillers des courtiers en assurance. Il fait l'inventaire des garanties et services des produits d'assurance du marché et permet de trouver la bonne offre rapidement avec l'argumentaire de vente associé. Ce système est proposé pour l'assurance automobile, les multi-risques habitations, la couverture d'emprunt, l'assurance incendie.

Gan Prévoyance, filiale du groupe **Groupama** s'appuie sur **DreamQuark** (2014, France, $3,5M) pour déployer une solution d'IA de deep learning afin d'améliorer la connaissance de ses clients et optimiser son développement commercial[589].

Riskgenius (2015, USA, $3M) est à l'origine d'une solution de traitement du langage qui gère le cycle de création et modification de contrats d'assurances. C'est une solution utile pour les agents d'assurance qui peuvent ainsi plus facilement comparer les polices d'assurances distribuées.

IBM Watson a été mis en œuvre dans **Insurance Assistant** de l'USAA (United Services Automobile Association), un agent conversationnel qui permet aux clients de cette assurance dédiée au personnel militaire US de s'y retrouver dans ses offres et services.

Une startup peut ambitionner de remplacer les compagnies d'assurance traditionnelles. C'est le cas de **Lemonade** (2015, USA, $180M), est une société d'assurance en ligne pour propriétaires et locataires basée à New York qui s'appuie fortement sur l'IA dans tous ses processus et se passer d'intermédiaires (les courtiers), y compris des chatbots dans la relation client (vidéo). La société utilise les sciences comportementales pour limiter la fraude. Par exemple, les constats sont réalisés par vidéo en ligne qui sont ensuite exploitées avec des solutions d'IA de détection d'émotion du style de celle d'**Affectiva** (2009, USA, $34,3M). Il y a aussi **Abe** (2017, USA), un courtier d'assurance incarné par un chatbot.

Natixis Assurances utilise enfin un système de RPA (robotic process automation) pour automatiser la cloture des contrats et la communication par email avec ses assurés.

Gestion des risques

Du côté de la gestion des risques, voici quelques offres et études de cas[590]. Elles portent sur l'évaluation des risques économiques, de cybersécurité et environnementaux pour l'immobilier.

Planck Re (Israël, $12M) fait de l'analyse de risques assuranciel exploitant des sources ouvertes sur les entreprises (images, textes, vidéos, sentiments, réseaux sociaux).

Cytora (2014, UK, $8,8M) réalise une analyse de risque et va jusqu'à aider à la tarification des assurances pour minimiser les pertes.

Cape Analytics (2014, USA, $31M) analyse les bâtiments par de l'imagerie aéricnne (donc avec force réseaux de neurones convolutionnels et de classification) permettant de qualifier la géométrie des bâtiments, la qualité des couvertures et celle de la végétation. Cela permet d'évaluer les risques dans l'immobilier. **Flyreel** (2016, USA) propose une offre similaire.

---

[589] DreamQuark développe des solutions d'intelligence artificielle à base de réseaux de neurones et de deep-learning avec des mécanismes d'auto-apprentissage capables d'explorer tous seuls tous types de données de les traiter. La startup propose des outils d'analyse via sa plateforme Brain qui permet d'explorer, optimiser et valoriser les données structurées (bases de données) et non-structurées (images, sons, voix) dans les secteurs de la finance et de l'assurance.

[590] Sur la gestion des risques dans les assurances, voir AI and risk management Innovating with confidence de Deloitte (32 pages), Impact of Artificial Intelligence on Reinsurance Sector Scor 2018 (36 pages) et The Rise Artificial Intelligence: Future Outlook and Emerging Risks, 2017 (24 pages).

Guidewire Software (2001, USA, $24,8M) a fait l'acquisition de **Cyence** (2014, USA, $40M) en octobre 2017. C'est une solution à base d'IA qui évalue les risques en matière de cybersécurité pour les entreprises clientes des assurances en exploitant de grandes masses de données ouvertes ou propriétaires tierces-parties. Cela permet aux assureurs de mieux adapter leurs offres aux clients entreprises et à minimiser les risques associés.

Côté assureurs, **Allianz**, exploite les données internes et externes à l'entreprise pour identifier la situation du client ou prospect. Il peut par exemple détecter que le client gare habituellement son véhicule dans une zone d'une ville où les vols sont plus nombreux que la moyenne et proposer une assurance contre le vol. Le tout est exploité dans un chatbot qui s'appuie sur IBM Watson.

PLANCK RE

insurance underwriting with SMB profiling using public web data: location, buildings, layout risk, management quality

Cytora

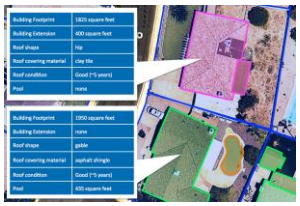
commercial insurers to target, select and price risk

achieves loss ratio improvement of up to 5% and reduce expenses by 10%

CAPE ANALYTICS

Flyreel

real estate imaging to assess environmental risk



GUIDEWIRE

assess cybersecurity risk with Cyence

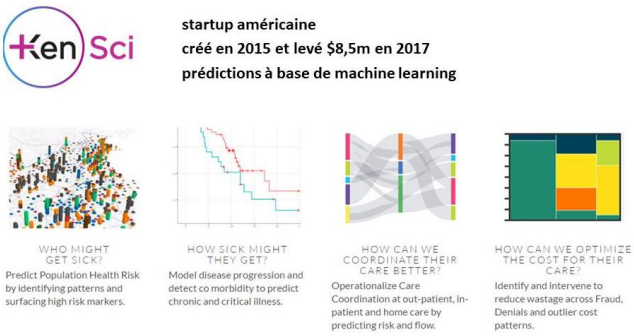
Les assurances sont partenaires de fournisseurs de solutions de maison intelligente pour réduire les risques de sinistres dans les logements ainsi qu'avec divers fournisseurs de solutions de santé pour réduire les coûts de santé, pour ce qui est des assureurs santé, surtout aux USA.

Dans l'assurance auto, on peut aussi encourager les conducteurs à faire auditer leur mode de conduite via des capteurs au standard CAN-2 dont l'offre est très abondante, notamment avec **Oocar** (2015, France).

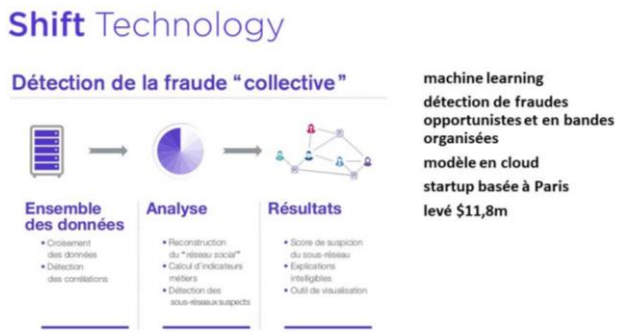
Gestion des sinistres

Element.ai (2016, Canada, $102M) est un éditeur de logiciels d'applications métiers qui s'appuient sur l'IA. L'un des marchés qu'ils gèrent est la gestion des sinistres dans l'assurance, pour la couverture des risques sur les biens physiques (immobilier, automobile). Ils gèrent la collecte d'informations structurées et non structurées sur les contrats, les clauses de couverture et les sinistres pour aider les assureurs à prendre position rapidement sur chaque dossier.

KenSci (2015, USA, $8,5M) est spécialisé dans les prévisions à base de machine learning. Il permet notamment aux assureurs santé de prévoir la structure de cout d'un parc d'assuré et de lancer des programmes de soin les diminuant (vidéo). Leurs modèles de machine learning permettent au passage de prédire l'espérance de vie de leurs clients et d'ajuster les « *dispositions à prendre pour optimiser leur fin de vie* ». Réjouissant[591] !



Shift Technology (2014, France, $39,8M) est spécialisé dans la détection de fraude organisée avec une solution en cloud. Elle permet par exemple d'identifier des redondances dans les scénarios et éléments fournis par les assurés, comme des photos qui seraient toujours les mêmes pour décrire les sinistres ! Ils sont déjà plus de 45 clients dont AG2R La Mondiale et démarrent leurs opérations à New York.



La détection de fraude à base d'IA est utilisée par l'assurance **Manulife** au Canada avec une solution développée en partenariat avec l'Université de Waterloo ainsi que par **Reliance Nippon Life Insurance** en Inde. Ce dernier évite ainsi de couvrir avec des assurances décès des personnes dont la mort prochaine est programmée à moins de trois ans !

Understory (2012, USA, $9,5M) est une startup dans les objets connectés qui fournit des capteurs d'environnement (humidité, température, vent, précipitations) assimilables à des stations météo miniatures qui permettent d'auditer a postériori l'origine de dégats d'origine météorologiques.

Tensorflight (2016, USA) utilise l'IA et l'imagerie aérienne pour automatiser l'inspection aérienne de biens immobiliers et évaluer d'éventuels dégats liés à des catastrophes naturelles[592].

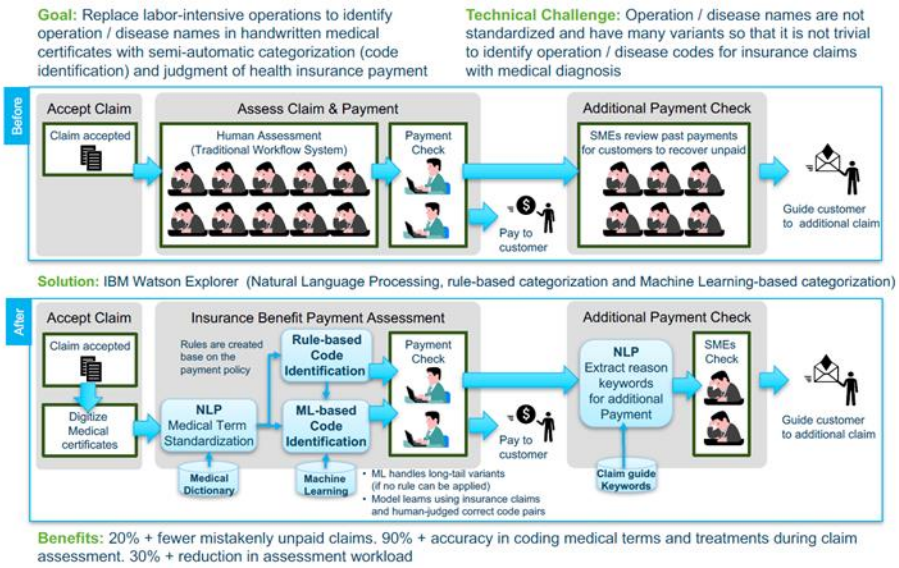
Les assurances font aussi appel à la reconnaissance d'image dans le cadre d'expertises, notamment automobiles ainsi que pour scanner les constats. **Tractable** (2014, UK, $34,9M) propose une solution d'inspection visuelle d'automobiles basée du deep learning pour de la classification automatique (vidéo).

Les processus internes aux assurances peuvent être automatisés avec des solutions et méthode de « Robotics Process Automation » déjà évoquée au niveau des services financiers. Elles peuvent par exemple faire appel à **Captricity** (2011, USA, $52M) et à sa solution de gestion documentaire.

[591] Voir KenSci research paper on End-of-Life Prediction to be presented at the AAAI 2018 conference, février 2018.

[592] Voir Roof Damage Assessment using Deep Learning, 2018 (6 pages) qui décrit les méthodes d'analyse de toits à base de deep learning, utiles pour les assurances en cas de catastrophe naturelle.

En 2017, l'assurance japonaise **Fukoku Mutual Life Insurance Co** faisait parler d'elle en annonçant prévoir de réduire de près de 30 % le personnel de son équipe d'évaluation des paiements grâce à une IA basée sur les briques logicielles d'IBM Watson et lancée en janvier 2017[593]. Soient 34 personnes (qui étaient en CDD de 5 ans) sur 131 (ce qui fait 26%...).



Cette IA exploite les divers documents médicaux pour gérer les remboursements de ses assurés santé en conformité avec le niveau de couverture.

Plus d'un an plus tard, il est difficile de savoir si la solution a tenu ses promesses. Il faut être toujours prudent avec ce genre d'annonce !

Toujours en 2017, un autre assureur japonais, **Nippon Life Insurance Co** mettait en production une AI d'analyse des meilleurs plans de couverture santé pour les clients grand public, exploitant les données de 40 millions de contrats.

Juridique

Le lancement de la startup **Ross Intelligence** (2014, Canada/USA, $13M) qui s'appuie sur IBM Watson il y a quelques années a créé un signal fort sur le marché : les métiers intellectuels comme celui d'avocat allaient être transformés radicalement par l'IA[594].

Vus de près, une bonne partie des outils de l'IA dans les métiers juridiques sont des moteurs de recherche améliorés qui permettent de consulter la jurisprudence et les lois. Des applications plus élaborées ont vu le jour pour produire plus rapidement des contrats, les contrats, les optimiser et pour générer des prévisions à bases probabilistes sur l'issue de procès[595].

Les techniques d'IA juridique tournent principalement autour du traitement du langage et de la modélisation des connaissances. Comme pour les chatbots, elles sont encore imparfaites car le langage et le raisonnement sont très difficiles à modéliser et manipuler par des machines. Ces solutions sont donc loin de remplacer les fonctions juridiques actuelles même si elles peuvent certainement améliorer la productivité des métiers du paralegal[596] dans les entreprises et cabinets d'avocats, en particulier aux USA qui sont les plus gros consommateurs, de loin, de cette profession.

---

[593] La source du schéma est Shaping the future of insurance with IBM Watson, novembre 2017 (43 slides).

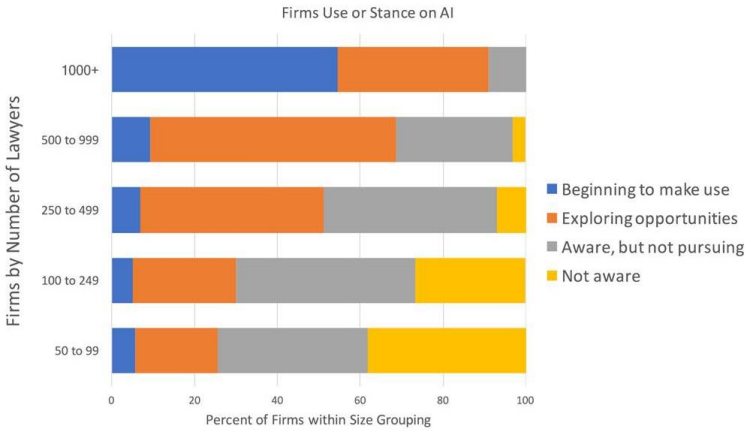
[594] Voir Legal Aspects of AI de Richard Kemp, novembre 2016, qui évoque à la fois les usages de l'IA dans les métiers juridique et les impacts juridiques de l'IA.

[595] Voir ce bon panorama dans The Robot Lawyer? de Saskia Mehlhorn, 2017 (41 slides).

[596] Voir l'excellent, long et très documenté Les robots, avocats et juges de demain ? Pas vraiment ... Intelligence artificielle en droit : derrière la "hype", la réalité par Emmanuel Barthe, mars 2018. D'une manière générale, les limitations des IA évoquées dans cet article sont liées au fait que la grande majorité des systèmes de recherche de texte juridiques exploitent des représentations mathématiques des textes dans des algorithmes de machine learning qui ne gèrent pas le sens des textes et n'ont pas de véritable système de représentation structuré des connaissances. Sauf si sont adjoints des moteurs de règles « à l'ancienne ».

L'IA juridique comprend des outils qui améliorent la productivité de nombreux professionnels, comme les tableurs l'ont fait pour tous les métiers exploitant des données chiffrées depuis plus de 35 ans. Comme toute technologie qui se déploie largement, l'IA pourrait aussi permettre un élargissement du marché juridique tout en ayant un effet déflationniste sur certains tarifs. C'est le phénomène bien connu de la commoditisation. L'étude de 2016 Altman Weil Law Firms in Transition 2017 indique que la moitié des cabinets d'avocats US de plus de 1000 salariés utilisaient déjà des outils d'IA. Ce taux est inférieur à 10% pour les autres tailles de cabinets.

Cela rappelle la situation de nombreuses professions libérales (notaires, experts comptables, médecins) qui sont fragilisées par leur fragmentation face aux ruptures technologiques qu'elles sont lentes à adopter vis-à-vis de grandes organisations plus structurées. Reste à savoir à quelles applications les personnes sondées faisaient allusion, surtout dans la mesure où la définition de l'IA est à forte géométrie variable pour la plupart des professionnels de tous métiers.



Dans **Legal Aspects of Artificial Intelligence**[597] de Richard Kemp (novembre 2016), on trouve cette petite liste intéressante d'étude de cas d'usage de l'IA dans des cabinets d'avocats américains. Il s'agissait à l'époque d'effets d'annonce, sans véritables retours d'expérience. Depuis, les choses ont avancé et l'adoption de ces outils a fait son chemin, surtout aux USA.

Table - Examples of recently announced B2B AI use cases in the legal services market

Date	Law firm	AI provider	Use case
2015			
Aug	Dentons	IBM/ ROSS Intelligence	Dentons partners with IBM on IBM Cloud. Dentons' NextLaw Labs partners with Ross Intelligence to develop a legal app powered by IBM Watson[23]
Sept	Berwin Leighton Paisner (BLP)	RAVN Systems	RAVN Systems announces that BLP is using its AI platform to manage property Light Obstruction Notices[24]
Oct		Thomson Reuters/ IBM Watson	Thomson Reuters partners with IBM to deliver Watson cognitive computing solutions[25], with Legal as the first use case
Dec	Riverview Law	CIXILEX	Riverview launches the Kim Virtual Assistant built on the CIXILEX platform acquired by Riverview in August.
2016			
May	Baker Hostetler	ROSS Intelligence	Baker Hostetler becomes the first US law firm to license ROSS

Date	Law firm	AI provider	Use case
May	BLP	not stated	BLP wins the first contested High Court application to use Predictive Coding in litigation document disclosure[26]
May	Linklaters	RAVN	Linklaters confirms it has signed an MSA with RAVN[27]
June	Allen & Overy	Deloitte	Allen & Overy launch digital derivatives compliance system MarginMatrix with Deloitte[28]
June	DLA Piper	Kira Systems	DLA Piper announces agreement to use Kira in M&A due diligence[29]
July	Clifford Chance	Kira Systems	Clifford Chance announces AI agreement with Kira Systems[30]
Sept	Freshfields	Kira Systems	Freshfields announces agreement to use Kira in its Legal Services Centre[31]
Sept	Slaughter and May	Luminance	Slaughter and May announces collaboration with Luminance on legal due diligence AI[32]

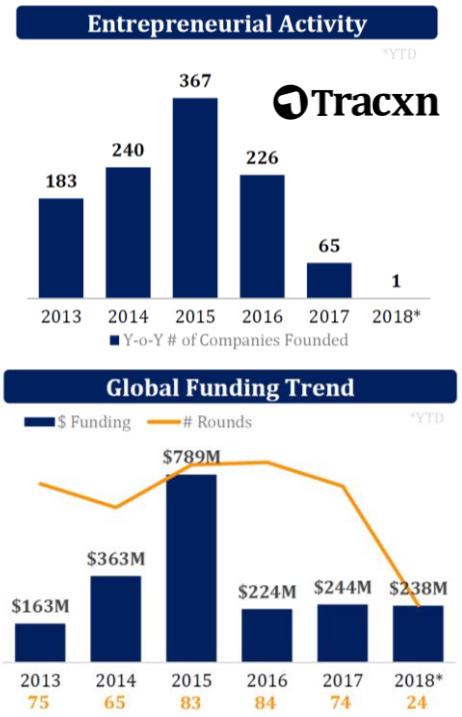
Comme pour tous les marchés traditionnels, celui des services juridiques est affecté par les innovations proposées par des startups. Elles sont relativement nombreuses, aussi bien aux USA qu'en France[598]. La cartographie *ci-dessous* en présente quelques-unes dont une partie seront détaillées dans la suite.

[597] Voir Legal Aspects of Artificial Intelligence, novembre 2016 (34 pages). Voir aussi L'intelligence artificielle au service de l'avocat : l'avocat-robot est-il à nos portes? d'Andrée-Anne Perras-Fortin et Eric Lavallée, septembre 2018, qui nous donne une perspective canadienne.

[598] Voir Legal tech report 2018 de la société d'analystes pour investisseurs en capital risque Tracxn, mai 2018 (214 pages) qui inventorie 1874 legaltechs dans le monde et cette Liste des legal techs françaises ou implantées en France de Benoît Charpentier datant de 2017 (6 pages) qui recense 94 startups legaltech en France, une partie seulement faisant appel à de l'IA. L'étude de Tracxn montre que le pic de création et de financement de legaltechs remonte à 2015, comme une indication que « les jeux sont faits » dans ce secteur.

Le Rapport 2018 de **Tracxn** fait ci-dessous un point intéressant des domaines où les startups des legaltechs ont le plus levé de fonds, sachant qu'une bonne part utilisent de plus en plus des briques d'IA ans leurs solutions.

La gestion de contrats arrive en premier suivi des outils de eDiscovery qui permettent de préparer un dossier juridique de contentieux et comprennent les moteurs de recherche juridiques. La propriété intellectuelle arrive en quatrième.



Key Sub Sectors in Legal Tech Sector

Sub Sectors	# Cos Tracked	# Funded	Total Funding
Legal Contract Management Companies offering contract lifecycle management platform which spans across contract...	240	70	$951M
eDiscovery Companies offering a platform to manage emails, documents, and media files for litigation...	130	52	$658M
Legal Documents/Forms Companies offering online DIY legal form services for individuals, SMBs, and Enterprises...	179	43	$377M
Intellectual Property (IP) Management Companies offering solutions for IP management including filing, tracking, analytics,...	174	59	$349M
Lawyer's Marketplace Companies offering legal services marketplace for connecting consumers, businesses, attorneys...	357	58	$203M
Law Enforcement Companies offering law enforcement software for law enforcement agencies	23	6	$130M
Legal Practice Management Companies offering legal practice management platform for end to end client intake to...	431	52	$101M

Moteurs de recherche juridiques

Une grande majorité des startups juridiques de l'IA proposent des moteurs de recherche dans les textes de lois et de jurisprudence.

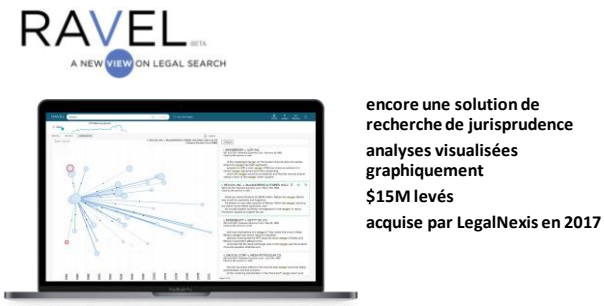
Elles fleurissent particulièrement bien dans les pays anglo-saxons dont le droit est influencé par la jurisprudence (dit de « case law »), tandis que le droit européen et surtout français, est plus fortement influencé par les lois et règlements (« civil law » ou droit romain).

Ross Intelligence (2014, USA/Canada, $13,1M), fondée par le canadien Andrew Arruda qui a fait une partie de ses études à la Sorbonne, s'appuie sur IBM Watson et est utilisé aussi bien par de grands cabinets d'avocats aux USA comme **Baker Hostetler** en mai 2016 ou petits comme **Salazar Jackson**[599]. C'est essentiellement un moteur de recherche que l'on interroge avec des questions posées en langage naturel. Il est censé devenir plus intelligent au gré de son usage.

Ce n'est pas évident à comprendre car la réponse à des questions ne constitue pas une forme d'apprentissage supervisé ou par renforcement. Ross est en fait spécialisé dans les affaires de propriété intellectuelle et dans la gestion des faillites d'entreprises.

Il propose aussi son outil EVA qui analyse des documents juridiques comme des briefs d'avocats de parties adverses pour les annoter automatiquement avec des éléments de droit et de jurisprudence, permettant d'évaluer rapidement la situation (vidéo).

Nous avons divers autres systèmes d'interrogation de bases de connaissances de jurisprudence comme **Ravel Law** (2012, USA, $15,2M, acquis par LexisNexis en 2017), issu des écoles de droit et d'informatique de Stanford (vidéo).



Casetext (2013, USA, $20,8M) qui est focalisé sur l'analyse de jurisprudences avec sa solution CARA (Care Analysis Research Assistant) (vidéo).

Elle indique notamment les lois qui ont été le plus souvent cassées en juridprudence, dans les pays de case law (vidéo).

Judicata (2012, USA, $7,8M) s'intéresse lui aussi à l'analyse de la jurisprudence. Ils détectent les propos déformés de manière délibérée par la partie adverse dans des procès.

Fastcase (1999, USA) propose un autre outil de recherche juridique (vidéo).

BloomBergLaw Points of Law est un moteur de recherche qui permet de préparer une plaidoirie en recherchant dans l'historique des opinions des juges. C'est encore une fois une solution adaptée uniquement au marché US.

Comme le droit est différent d'un pays à l'autre, les startups du secteur sont souvent cloisonnées au départ par pays. Très peu de startups couvrent à la fois les marchés juridiques des deux côtés de l'Atlantique.

[599] Voir Law Firm BakerHostetler Hires A 'Digital Attorney' Named ROSS, mai 2016 et Legal AI: It's not just for Big Law – Salazar Jackson and ROSS Intelligence, janvier 2017.

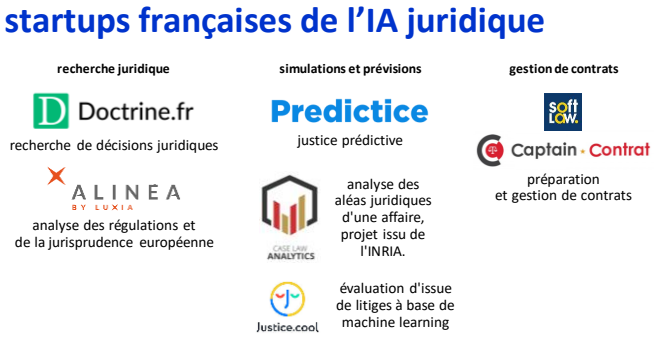
Il y a aussi **Doctrine.fr** (2016, France, $12M)[600] qui permet de faire des recherches en langage naturel dans une grande masse de bases de jurisprudence pour 129€ HT par mois (vidéo). Sachant qu'il n'y a pas beaucoup d'IA dans la solution. Elle sembler reposer en grande partie sur l'interrogation classique de bases de données.

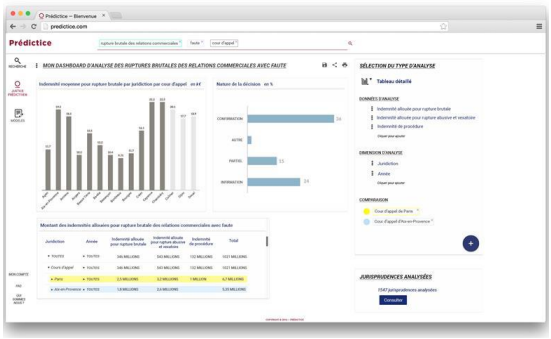
Simulations et prévisions

Les outils de simulation et de prévision, souvent assimilés à la notion de justice prédictive, analysent des situations juridiques pour évaluer les risques et leur issue. Ils ne remplacent ni les avocats ni les juges même si parfois, ils apportent un peu plus de rationnalité que ces derniers. Toutefois, cette rationnalité est à prendre avec des pincettes puisqu'elle s'appuie sur la jurisprudence qui est elle-même d'origine humaine ! Ceci étant, la perspective d'une justice prédictive exploitée par les juges eux-mêmes commence à effrayer. Avant même que cela fonctionne, elle inquiète déjà la profession et les colloques de juristes sur le sujet sont nombreux depuis deux ans.

Case Law Analytics (2017, France) est une startup proposant une solution d'analyse des aléas juridiques d'une affaire. Là encore, il doit y avoir beaucoup de traitement du langage au programme. L'un des cofondateurs est un ancien directeur de recherche de l'INRIA spécialisé en modélisation mathématique en économie, finance et droit.

La solution analyse la jurisprudence, établit et visualise des modèles probabilistes permettant d'évaluer non seulement l'issue d'une affaire mais également ses éléments quantitatifs comme les dommages et intérêts. Comme partout dans l'IA, la qualité d'une évaluation dépend de la quantité d'études de cas qui entraînent les modèles d'IA. Si votre cas est inédit, l'étude de la jurisprudence n'apportera pas grand-chose. Et elle est pertinente si la jurisprudence est abondante et bien commentée. Depuis fin 2017, le cabinet Flichy Granger spécialisé en droit social est équipé de Case Law Analytics[601].





Predictice (2016, France) est aussi positionné dans la justice prédictive. C'est cependant une solution généraliste avec un moteur de recherche de documents juridiques. La startup exploite les données ouvertes de Légifrance (textes de droit) et Jurica (jurisprudence).

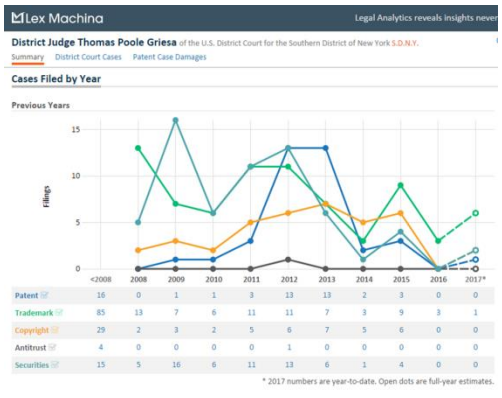
En matière de prévisions, le système fournit surtout une vue statistique et probabiliste des cas passés et les moyennes de dommages et intérêts. Les outils exploités intègrent l'analyse de langage et un moteur de règles.

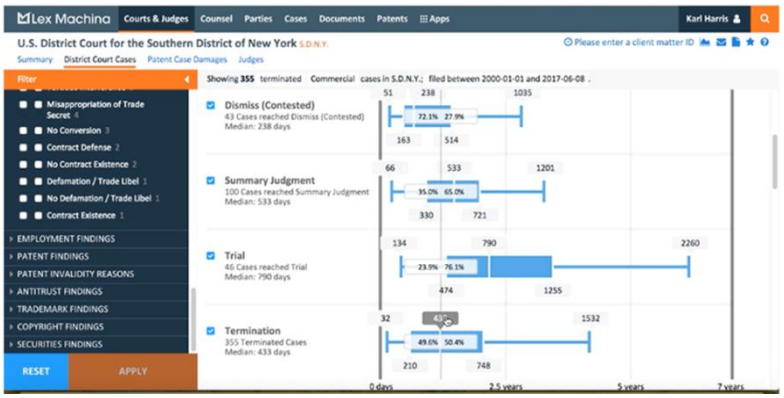
---

[600] Voir Entre levée de fonds record et levée de bouclier de la profession, quel avenir pour Doctrine ? d'Anais Richardin, juin 2018, qui décrit les méthodes de growth hacking contestées de Doctrine.fr. La startup a été attaquée en justice par l'ordre des avocats du bareau de Paris en septembre 2018 pour usurpation d'identité, la méthode qu'aurait employé la startup pour récupérer des informations sur Infogreffe. Voir La start-up Doctrine attaquée par l'Ordre des avocats de Paris de Guillaume Bregeras, paru dans Les Echos le 27 septembre 2018.

[601] Voir Flichy Grangé Avocats utilise Case Law Analytics, un outil de justice prédictive, novembre 2017.

LexMachina (2009, USA, $10M, acquis par LexisNexis en 2015) fournit des outils d'analyse statistiques pour les avocats. Ces analyses sont plutôt quantitatives et temporelles. Elles permettent d'anticiper la longueur des procédures, la réaction des avocats de parties adverses et des juges et se s'organiser en conséquence. C'est une sorte de « Business Objects » pour avocats.

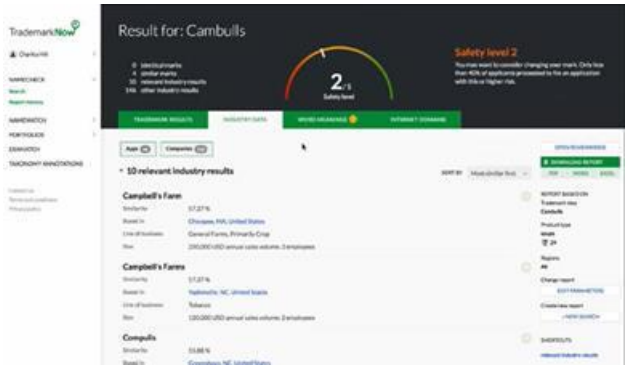




Propriété intellectuelle

La propriété intellectuelle est un champ à part entière du droit avec ses spécialistes, les conseils en propriété intellectuelle. Les bases de données associées sont spécifiques, à savoir celles des marques et des brevets.

- **Juristat** (2012, USA, $1,8M) réalise des analytics sur les données publiques sur les brevets et les avis des agents de l'USPTO pour optimiser les stratégies de protection de brevets. C'est une solution dédiée au marché US. Le financement de la startup l'explique en partie. Il est encore limité pour permettre un développement international rapide.
- **Turbo Patent** (2008, USA, $6,7M) produit des reportings sur la qualité et la valeur d'un portefeuille de brevets.
- **Lex Machina** (2009, USA, $10M, acquis par LexisNexis) fournit une solution de prévision sur les litiges de propriété intellectuelle. Elle exploite des solutions d'analyse du langage développées à l'Université de Stanford.
- **Data&Data** (2012, France) détecte en temps reel les ventes de contrefaçons et de marché gris sur Internet. L'outil à base d'IA s'appuie sur la détection d'anomalies dans les prix et les images des articles en vente.
- **TrademarkNow** (2012, USA, $9,4M) propose une solution de recherche portant sur la protection des marques (*ci-dessous*). Il est cependant difficile d'y identifier des morceaux d'IA.

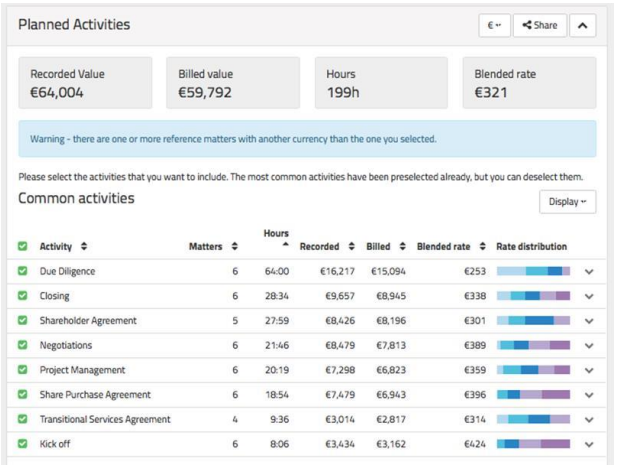


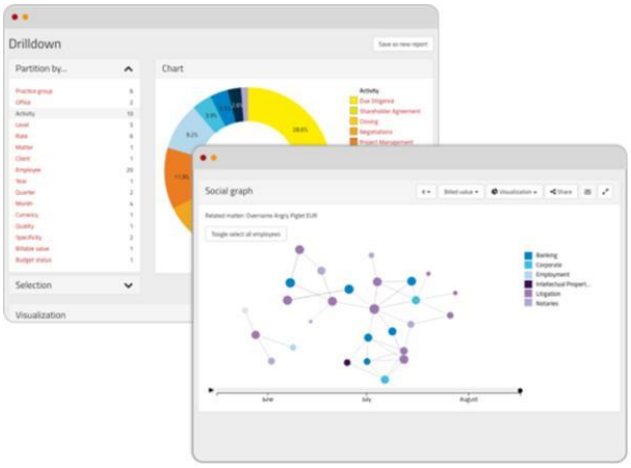
- **Clarivate Analytics** (2016, USA) propose une large gamme d'outils de recherche dans la propriété intellectuelle couvrant brevets et marques, dans ce dernier cas, via l'acquisition de TrademarkVision qui utilise le deep learning pour reconnaître les logos de marques et faire des recherches d'antériorité, ainsi qu'une solution de lutte contre la contrefaçon issue de l'acquisition de MarkMonitor.

Gestion de contrats

Nombre d'applications juridiques d'entreprises sont destinées à faciliter la gestion de contrats, pour identifier les clauses clés ou anormales, gérer les versions et faciliter le circuit de préparation et de signature. Les contrats commerciaux ou d'associations entre entreprises constituent en effet une grosse part du volume de travail des services juridiques des entreprises.

ClockTimizer (2014, Pays-Bas) propose un outil de business analytics pour cabinets d'avocats qui permet d'évaluer le temps passé sur des contrats clients et d'affiner ensuite les devis pour d'autres clients basés sur l'expérience. Les outils exploitent à la fois des données textuelles (mots clés des contrats, etc) et quantitatives (temps passé, etc).





Kira Systems (2015, Canada, $50M) propose une solution de « due diligence », de recherche, d'analyse et d'édition de contrats. Elle gère notamment les contrats de fusions-acquisitions ou de restructuration de capital d'entreprises.

C'est aussi l'activité de **eBrevia** (2012, Canada, $4,3M) qui a codéveloppé sa solution avec l'Université de Columbia qui couvre notamment le droit de l'immobilier et fournit des solutions d'analytics de contrats.

LawGeex (2014, Israël, $21,5M) fait de la revue de contrats, notamment de confidentialité (NDA). Une étude montre que LawGeex est plus efficace qu'un avocat moyen dans l'évaluation de contrats de NDAs. En février 2018, leur solution obtenait un score de 94% de détection d'anomalies dans des NDAs contre 85% pour une vingtaine d'avocats humains[602].

[602] Voir Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts (37 pages) et LawGeex Hits 94% Accuracy in NDA Review vs 85% for Human Lawyers, février 2018.

Hyperlex (2017, France, 1M€) est une plateforme en ligne de gestion de contrats. Elle analyse les contrats d'entreprise pour identifier les clauses clés et permet leur revue collaborative dans l'entreprise. L'outil comprend surtout un dashboard de suivi de contrats et de leur circuit de signature, une fonction de workflow assez traditionnelle.

Neota Logic (2010, USA) est spécialisé dans la préparation de contrats de confidentialité (NDAs). L'IA qu'elle contient est censée permettre la sélection des bons templates en fonction des besoins. Ca peut être aussi bien un moteur de règles pas trop complexe ou un outil de machine learning exploitant quelques dizaines de variables et de la PCA (Principal Components Analysis) pour identifier les paramètres clés de choix des templates.

Klarity (2017, USA, $120K) gère la revue de contrats en cloud dans des fichiers Word qui sont automatiquement annotés.

Legal Sifter (2013, USA, $6,3M) est une autre solution d'analyse rapide de contrats.

Autres usages

Au programme, nous avons notamment des outils de business analytics et des chatbots grand public.

On compte notamment **LegalZoom** (1999, USA), un service d'avocat en ligne couvrant à la fois le droit des affaires et le droit civil[603] qui s'appuie sur IBM Watson.

Quid-IA (2018, France) propose des chatbots juridiques pour les entreprises et les particuliers. Ce serait le premier chatbot accompagnant les entreprises dans leur mise en conformité au RGPD. C'est en fait une filiale du cabinet juridique ALTIJ de Toulouse.

Cognigo (2016, Israël, $10,9M) aide également à gérer sa conformité au RGPD avec sa solution DataSense qui analyse de nombreux flux de données de l'entreprise. La cuisine interne de l'ensemble n'est cependant pas bien claire. Ils indiquent surtout exploiter le traitement du langage pour identifier le type d'information qui circule, issue de données structurées et non structurées. Ils ont aussi l'air d'analyser la cybersécurité des données personnelles, notamment celles qui sont gérées dans le cloud.

DoNotPay (2016, UK) est un chatbot anglais créé par un jeune qui avait à l'époque 19 ans et qui permet de faire sauter les contraventions aux USA et au Royaume Uni et pour négocier la baisse de vos billets d'avions achetés, lorsque les prix baissent (on demande à voir…). Il a ensuite été étendu à la gestion de nombreux cas de droit civique. L'ensemble exploite IBM Watson, ce qui montre qu'avec un peu de détemination une personne isolée peut le paramétrer efficacement.

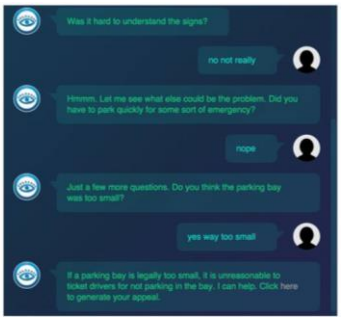
DoNotPay

chatbot UK lancé par un jeune de 19 ans

site gratuit

permet de faire sauter lescontraventions

efficace dans 65% des cas



Au passage, signalons que nombre de logiciels, les inventions à base d'IA sont brevetables si elles comportent des spécificités techniques nouvelles. Se pose la question de la brevetabilité d'une invention qui serait créée par une IA. On n'y est pas encore car la plupart des IA impliquées dans des inventions sont des outils exploités par des inventeurs homo-sapiens. Aucune IA ne gère une invention de la tête aux pieds, de l'idée ou intention jusqu'à la formalisation et la mise en œuvre. C'est toujours un outil[604].

[603] Voir le numéro d'août 2014 de The American Lawyer qui en parle bien.

[604] Voir Les robots, des inventeurs comme les autres ? de Magali Touroude, avril 2018.

AirHelp (2013, USA, $12,3M) a développé un « robot-avocat » dénommé Laraqui détermine les chances d'obtenir gain de cause en justice en fonction des éléments clés du dossier. Cela concerne en fait les réclamations dans le transport aérien.

Une autre startup, **RefundMyTicket** (2014, France) est aussi positionnée sur le créneau des remboursements de vols annulés ou retardés dans le transport aérien. Le site **Justice.cool** au sein du groupe **Claim Assistance** qui comprend Refundmyticket, évalue la légitimité des demandes par similarités avec un mix de machine learning et de moteur de règles.

Ravn Systems (2010, UK, acquise par iManage en 2017) a développé l'application LPP (Legal Professional Privilege) pour le Serious Fraud Office, l'organisme de l'état Britannique qui gère les grandes affaires de criminalité financière, fraudes et corruption, une sorte de Tracfin étendu. LPP leur sert à passer au peigne fin les documents juridiques des affaires en cours.

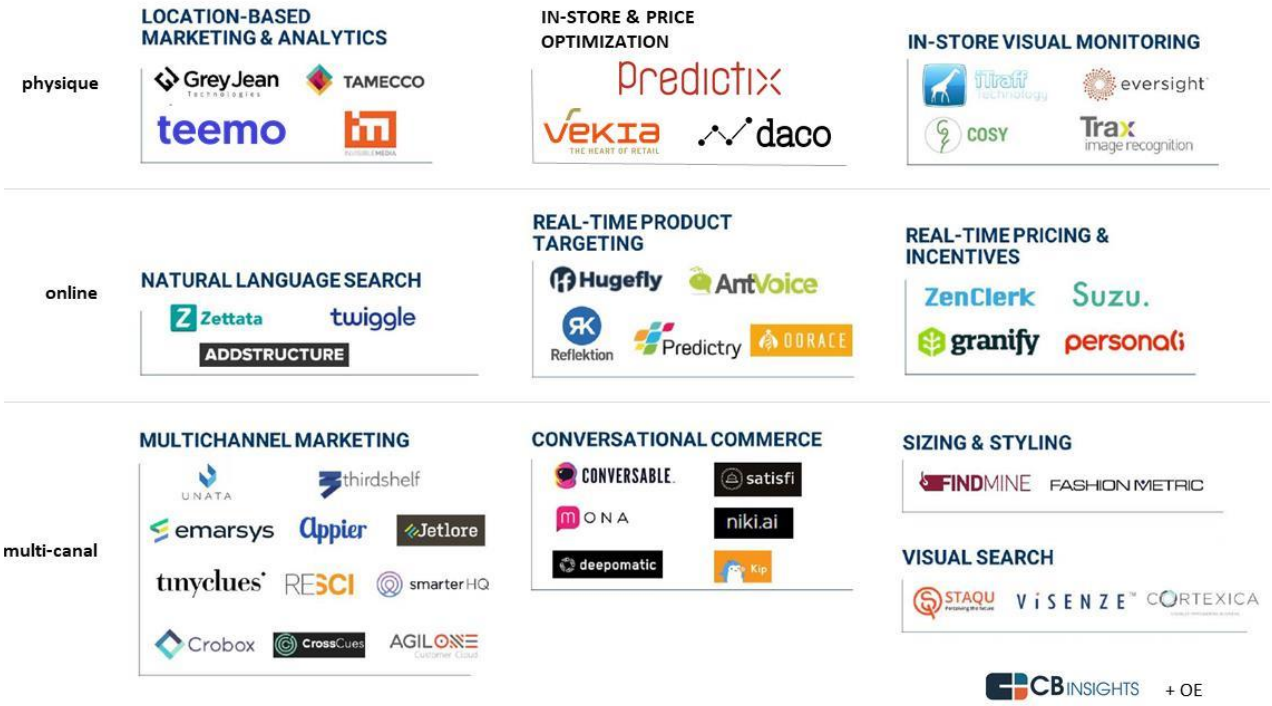
Distribution

Le monde de la distribution est un autre excellent terrain pour déployer des applications à base d'IA[605]. Les contrastes sont marquants entre le commerce en ligne et le commerce traditionnel. Les premiers sont des utilisateurs intensifs de numérique à tous les étages, y compris d'IA et les seconds ont des processus plus traditionnels qui évoluent bien plus lentement, malgré les tombereaux de données qu'ils sont en mesure d'accumuler. C'est même le cas chez les grandes enseignes qui ont pignon à la fois sur rue et sur Internet.

Les innovateurs du secteur doivent conserver dans leur mire les clients. Leurs besoins ne sont pas sorciers : ils souhaitent trouver rapidement ce qu'ils cherchent, pouvoir comparer les offres et trouver les meilleures au niveau fonctionnel et tarifaire, pouvoir obtenir le produit rapidement, et pouvoir le dépanner ou retourner facilement en cas de problème. La distribution de détail a aussi besoin de rendre ses points de vente attirant visuellement, émotionnellement et dans la qualité de service, pour éviter l'inexorabilité de la fuite vers le commerce en ligne.

Il ne faut jamais perdre la dimension humaine ! Mais attention ! L'IA ne doit pas être trop porteuse d'espoirs déplacés pour ce secteur d'activité en crise face à des bouleversements sociétaux sur les modes de consommation, de transports, de relation au temps et de relations interpersonnelles ? L'IA ne doit pas relever de la méthode Coué servant à éviter la vue désagréable des crises à venir pour le retail physique.

[605] Voici quelques sources d'information où j'ai pu glaner quelques éléments sur l'IA dans le retail : Artificial intelligence for High Frequency Retail – Pricing, Inventory and Margins Optimization, septembre 2018, The retail renaissance: Leading brands use data and AI to win, avril 2018, Artificial Intelligence in Retail, Part 1: Applications Across Customer-Facing Functions de Coresight Research, 2018 (12 pages), Artificial Intelligence In The Retail Industry de Shaily Kumar, novembre 2017 et AI retail playbook de Microsoft (29 slides).



Les besoins des commercants ? Ce sont des intermédiaires entre les marques et les consommateurs. Ils doivent analyser les tendances, comprendre les sentiments autour des marques, mettre les produits bien marketés dans les maisons des consommateurs, faire de l'upselling et du cross-selling, réduire leurs frais de gestion, optimiser les stocks et leur rotation, limiter la fraude (en ligne) et la démarque inconnue (en magasin). En gros, les retailers veulent soit prédire le futur, soit l'influencer à leur avantage.

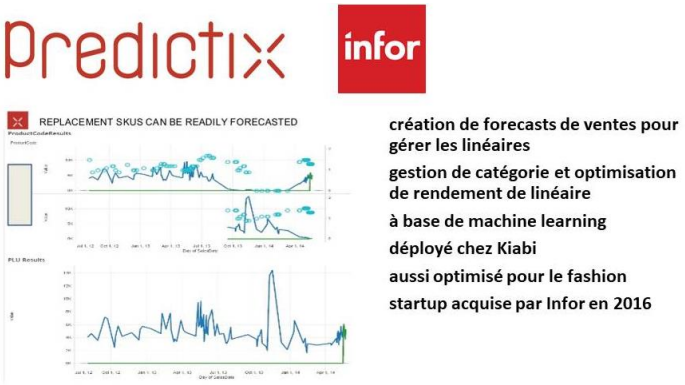
Nous allons ici faire le tour de quelques startups et études de cas d'usages de l'IA dans le retail en séparant si besoin est le retail physique du commerce en ligne[606].

IA pour le retail physique

Voici quelques startups qui exploitent diverses briques d'IA pour répondre aux besoins de retailers traditionnels côté assortiment produit, optimisation des rayons, lutte contre la démarque inconnue, dans le web-to-store et la recommandation.

[606] Dans Why retail's artificial intelligence bet is all wrong, mars 2018, Luis Perez-Breva calme les ardeurs du "tout IA" dans le retail en mettant en avant le fait que l'IA doit aider les humains, pas les remplacer, notamment dans les points de vente. Il rappelle aussi que les retailers disposent de moins de données que les e-commerçants sur le profil des acheteurs, tout du moins dans la distribution spécialisée. En France, un hypermarché sait beaucoup de choses sur vous, même s'il n'exploite pas bien ces données. Il avance aussi que la recommandation est survendue, y compris dans le cas d'Amazon. Il la juge faiblement pertinente mais attractive pour les clients car cela les occupe. Il relève aussi un point clé que je met en avant depuis quelques années : le fait que le logiciel et l'IA transfère du travail rémunéré des fournisseurs vers du travail non rémunéré chez les clients.

Predictix (2005, USA, $40M) est une startup créée par des Tunisiens. Elle fait partie du groupe américain Infor depuis 2016. C'est un spécialiste de l'optimisation de linéaires. Comme son nom l'indique, elle doit utiliser des techniques de machine learning pour faire du prédictif. **Celect** (2013, USA, $15,2M) permet aussi d'optimiser l'assortiment des rayons en fonction d'analyses prédictives comportementales des clients.



Vekia (2008, France, $14,9M) utilise le machine learning pour faciliter l'optimisation d'approvisionnement (vidéo). Sa solution est déployée chez Leroy Merlin, But, MrBricolage et Jacadi.

Reflektion (2012, USA, $42,8M) aide les commerçant à convertir les prospects en clients à partir de son moteur de recherche de point de vente pour les grandes enseignes comme Disney.

Teemo (2014, France, $17,9M), anciennement Databerries, est un spécialiste du ciblage de clients mobiles. Il les cible en fonction des lieux qu'ils visitent, par triangulation des signaux Wi-Fi, si celui-ci est activé dans les smartphones.

Ils ont déjà plus d'une centaine de clients dont Leroy Merlin, Jardiland, Carrefour et Casino.

Occi (2015, France) permet aux commercants d'envoyer des recommandations personnalisées à leurs clients, et sur leur smartphone, pour peu qu'ils l'aient convaincu d'installer son application ou qu'ils disposent de son numéro de téléphone ou de ses identifiants de réseaux sociaux.

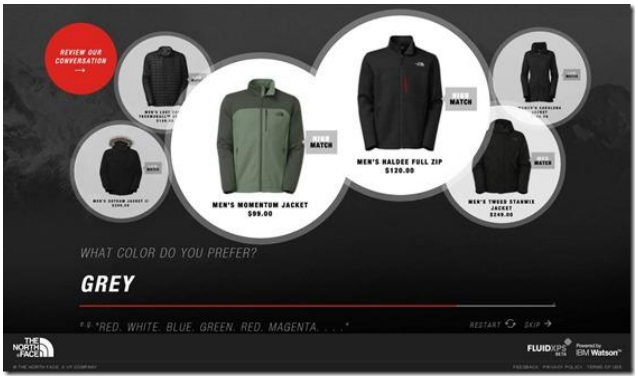
Percolata (2012, USA, $14,7M) exploite caméras de surveillance, captation audio, détection de smartphones et machine learning pour prédire le trafic dans les magasins. Il croise ces données avec l'historique de performance des vendeurs pour planifier les équipes de vente générant le plus haut niveau de chiffre d'affaire.

Angus.AI (2014, France, $500K) propose une solution de gestion des linéaires dans la distribution qui s'appuie sur des caméras de surveillance et celles des smartphones. Elle permet de gérer plus efficacement le réassortiment des rayons et le contrôle des prix.

La solutions serait évaluée ou déployée chez Carrefour et Leroy Merlin[607].

Quidivi (2006, France, $1,4M) est une startup française qui analyse le visitorat en magasin via caméras et machine vision. Ses outils mesurent non seulement le trafic mais aussi l'attention. Ils détectent l'âge et le sexe des visiteurs. La startup est ancienne dans l'écosystème. Elle a fait évoluer ses techniques d'analyse d'image au gré des évolutions technologiques du deep learning. Le suédois **Modcam** est un concurrent.

IBM propose une solution d'analyse des données clients et de sources diverses pour anticiper les besoins du marché et adapter les inventaires et les stratégies de tarification. IBM propose aussi un **Personal Shopper** été réalisé en partenariat avec **Fluid** (1999, USA, $24M, acquis par Astound Commerce en 2018). Le premier client est la chaîne de distribution de vêtements sportifs **North Face**. Il s'agit là encore d'un agent conversationnel utilisable via le service en ligne du site marchand.



Le corpus de données utilisé exploite tout le catalogue du site ainsi que les différents critères de choix des vêtements. Le dialogue proposé est très "scripté". Son arborescence semble limitée. Le système a été présenté au Big Show 2016 de la National Retail Foundation à New York[608].

Lokad (2008, France) est un spécialiste de la supply chain pour le retail. Ils aident les commerçants à prévoir ce qui va leur arriver, notamment la demande client. Ce sont de gros utilisateurs de machine learning.

RoofStreet (2015, France, 1M€) analyse et anticipe les déplacements des personnes pour leur faire ensuite des propositions adaptées d'implantations de points de vente et de l'analyse de concurrence. On appelle cela le géomarketing prédictif.

Twenty Billion Neurons (2015, Canada, $12,5M) analyse de son côté des déplacements des clients dans les points de vente.

UntieNots (2015, France) a développé pour Auchan un défi personnalisé Waaoh pour gagner je ne sais quoi aux clients. Le défi est généré à partir de machine learning sur l'historique des achats et dépenses. Je suis client d'Auchan depuis des décennies et utilisateur de leur carte de fidélité mais je n'ai rien vu. Ce sont les affres du marketing !

Diverses solutions apparaissent pour automatiser le check-out et le paiement dans le retail, comme avec **Aipoly** (2015, USA) qui utilise des caméras pour détecter les produits achetés (vidéo). Je les avais découverts au CES 2018.

---

[607] La startup avait été créée par des anciens ingénieurs d'Aldebaran qui ont développé la partie logicielle des robots Nao et Pepper. Leur première offre était une solution logicielle embarquée dans les robots leur apportant les fonctions de base de reconnaissance vocale et faciale et de détection d'obstacles. Le tout fourni sous la forme d'un SDK et d'APIs (interfaces de programmation). Ils s'appuient beaucoup sur des solutions open source du marché. Le marché du retail leur a ensuite semblé plus prometteur, et hop, un pivot !

[608] Pour en savoir plus voir ce compte-rendu détaillé sur le JDN : Comment The North Face a appliqué Watson à l'expérience d'achat, de Flore Fauconnier, janvier 2016.

Rubikloud (2013, Canada, $45,5M) promet monts et merveilles aux retailers. En gros, vous alimentez leur système avec toutes vos données, et il va vous proposer les promotions à lancer pour optimiser les ventes, prévoir vos ventes et les augmenter de 10% et les résultats de vos campagnes marketing. C'est mis en œuvre à Hong Kong, donc un peu loin pour vérifier ce qu'il en est !

Everseen (2007, Irlande, $13,5M) a créé timi.ai, qui détecte la fraude en sortie de caisse dans les points de vente via l'usage de caméras de surveillance.

StopLift (USA) est aussi utilisé pour détecter la fraude mais pour les caisses automatiques self-service. Toujours avec usage de caméras de surveillance pour vérifier que les produits sont bien tous scannés et pesés.

Il faudrait ajouter ici les outils de robotisation des entrepôts de la distribution, un domaine où les e-commerçants sont généralement plus avancés du fait d'une chaîne logistique plus sophistiquée b2c alors que la chaîne logistique des commerçants physique est plutôt b2b, allant de stocks en stocks et de stocks en rayons. Ce sujet est traité dans la partie logistique de cette rubrique.

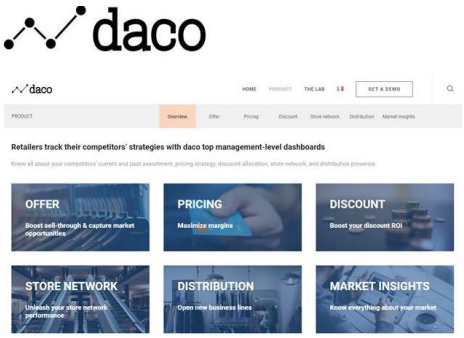
IA pour le commerce en ligne

Le commerce en ligne intègre évidemment presque tout ce que l'on trouve de nouveau dans le marketing, pour le ciblage publicitaire et commercial, sur les techniques d'upselling et cross-selling basées sur la recommandation qui s'appuient sur le machine learning ou dans la relation clients, chatbots compris.

Le commerce en ligne peut exploiter quelques autres nouveautés à base d'IA :

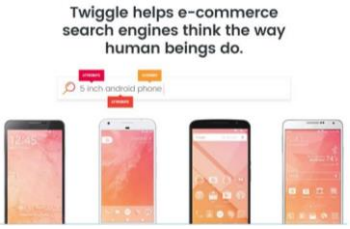
- **Commerciaux sédentaires** : l'optimisation de leur activité et la prévision du comportement des clients avec des startups particulièrement bien financées est proposée par **InsideSales** (2004, USA, $251M), **Clari** (2012, USA, $61M), **Wise.io** (2012, USA, $3,6M, acquis par GE en 2016) et **Spiro** (2014, USA, $4,5M). Il est assez difficile de départager toutes ces belles startups !
- **Optimisation du parcours client** en ligne avec **Gainsight** (2011, USA, $156M), **Jetlore** (2011, $10,6M, acquis par Paypal en 2018) et **OnCorps** (2011, USA, $2,3M). **Granify** (2011, Canada, $13M) va jusqu'à suivre pas à pas le parcours en ligne des clients pour détecter ceux qui pourraient abandonner le panier en cours de constitution et leur proposer une action ou information permettant de l'éviter. **Influans** (2016, France, 6M€) optimise aussi le parcours client pour leur proposer le bon produit avec les bonnes incitations et au bon moment. **TargetToSell** (2012, France, 5M€) s'appuie sur un mix de machine learning classique et d'un réseau de neurones pour optimiser le parcours du parcours en fonction du profil du visiteur et de ses goûts de produits captés au niveau du site, cela va jusqu'à personnaliser le bandeau du site en fonction du profil de l'utilisateur. **Nextuser** (2013, USA, $2,3M) se positionne comme un outil d'intégration des outils marketing. Enfin, **Perfect Path** (Canada) propose aussi un outil de modélisation du parcours client qui sert à définir les meilleurs scénarios de transformation en fonction des objectifs et segments visés.
- **Solutions pour trouver la bonne taille** et le bon style pour s'habiller avec **Thread** (2012, UK, $16,32M), le styliste en ligne **Stitch Fix** (2011, USA, $79,4M) qui exploite du dialogue en langage naturel, **Volumental** (2012, Suède, $4,5M) pour le choix de ses chaussures et **Thirdlove** (2013, USA, $13,6M) pour la taille de son soutien-gorge.
- **Optimisation du pricing**, comme avec **Brennus Analytics** (2015, France) qui est focalisé sur les ventes b2b et peut ajuster les prix de manière dynamique avec des techniques de machine learning multicritères. Leur solution analyse, compare et prédit les comportements d'achat des clients pour adapter les prix à la stratégie marketing et commerciale comme les gains de parts de marché, de marge tout en prenant en compte les contraintes comme la charge des usines, des stocks.

- **Benchmark des prix et des promotions** des concurrents, comme avec **Daco** (2016, France) qui s'appuie sur du deep learning et de la reconnaissance d'images et fait du prédictif sur les actions des concurrents, le tout avec une couverture mondiale. Ils ont notamment comme clients L'Oréal, Vente-privée et les Galeries Fafayette.



analyse des ventes
dataviz
optimisation de prix
IA non précisée !
startup française
crée en 2016

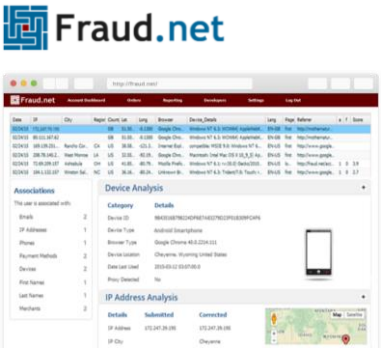
- **Moteurs de recherche** avec l'israélien **Twiggle** (2014, Israël, $35M) qui permet de faire des recherches textuelles et imite le comportement d'un commercial traditionnel et puis, dans les moteurs de recherche d'images pour optimiser la gestion de site de vente en ligne avec **ViSenze** (2012, Singapour, $14M), **Cortexica** (2008, UK, $9,2M) et son logiciel findSimilar en cloud, **Grokstyle** (2016, USA, $2,5M) et **Slyce** (2012, USA, $37M de levés, IPO en avril 2015). Et puis le moteur de recherche FashionBot de **GoFind** (2016, USA) qui permet de retrouver dans un site en ligne ce que l'on trouve dans un magasin.



moteur de recherche pour applications de ecommerce
techniques de NLP appliquées à la recherche de produits
API en cloud
startup isrélienne créée en 2014
$35m levés

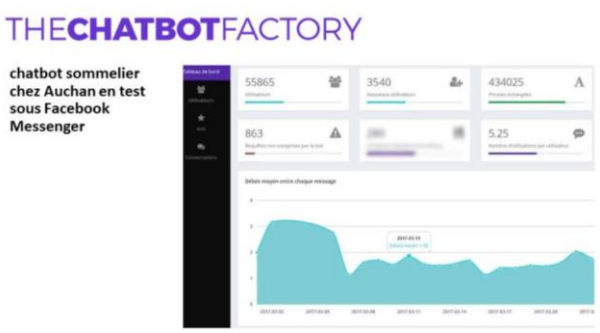
- **Systèmes de recommandation** basé sur du machine learning avec **Turi** (2013, $23,5M, acquis par Apple en 2016), une startup a été montée par des anciens de Carnegie Mellon sous la forme initiale d'un projet open source. Même histoire avec **Reflektion** (2012, USA, $42,8M), adopté par Disney et Converse, qui propose du ciblage produit temps réel. Et aussi **Kameleoon** (2008, France, 2,3M€) qui fait notamment de la personnalisation et de la recommandation ciblée aux visiteurs d'un site web à base de machine learning. La solution est déployée chez Cdiscount et Allopneux. Enfin, **Skapánê** (2015, France, 500K€) propose diverses solutions exploitant le machine learning pour faire des prévisions en temps réel pour la recommandation de produits, la maintenance prédictive et la lutte contre la fraude. Basés à Lille, ils ont notamment pour client Auchan, ce qui se comprend.

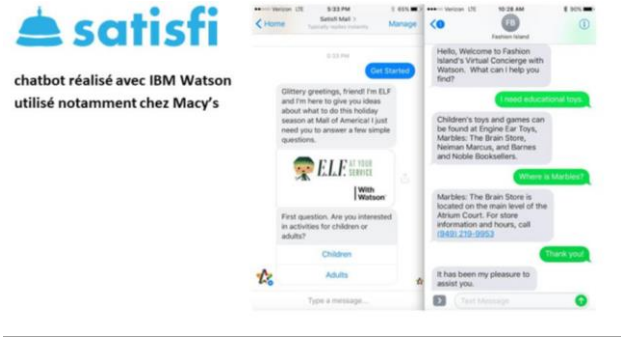
- **Détection de fraude** pour le commerce en ligne avec **Fraud.net** (2015, USA) qui s'appuie probablement sur du machine learning.
- **Détection de contrefaçons** chez **Data & Data** (2012, France) qui analyse les ventes de produits contrefaits et marchés gris sur Internet (sites de e-commerce, places de marché, réseaux sociaux comme Facebook et Instagram). Cela permet notamment de vérifier la fidélité de son réseau de revendeurs autorisés.



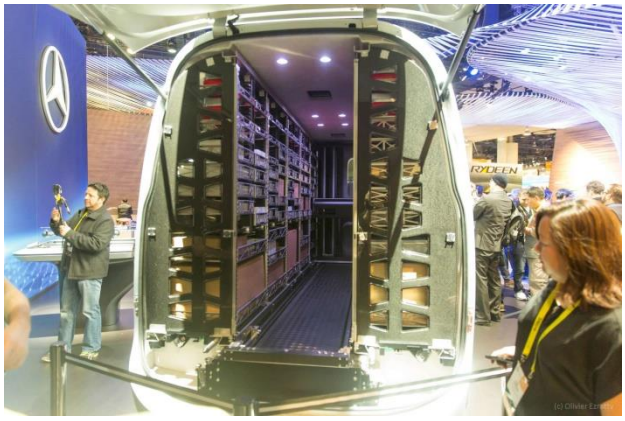
agrège les données sur les fraudes de site de ecommerce en temps réel
protège 2% des sites US
détecte 100 modèles de fraude par jour
USA, 2015

- **Chatbots** en tout genre avec l'américain **Satisfi Labs** (2016, USA, $1M) et l'agence **The-ChatbotFactory** (2012, France) qui créé des chatbots sur mesure. Elle a créé un chatbot sommelier pour Auchan sous Facebook Messenger. **La Redoute** a déployé début 2018 un agent vocal dans son application mobile qui est associé à un outil de reconnaissance visuelle de produits photographiés avec son smartphone.





Les e-commerçants font aussi appel à des solutions d'automatisation de la gestion des entrepots pour peu qu'ils aient la taille critique.



Les livraisons sont souvent réalisées par des sous-traitants spécialisés (ColisPrivé, …). Ils pourront un jour mettre en route la camionnette à propulsion électrique robotisée de **Mercedez-Benz** dont l'intérieur comprend un robot de manipulation de colis qui les transmet à deux drones via des ouvertures sur le toit. Le « last mile » par les airs !

Les drones viennent de **Matternet** dans lequel Mercedes a investi 562M€[609]. Et Amazon est le premier client en vue pour ce genre de solution. C'était un prototype présenté au CES 2017 mais il ne semble pas encore opérationnel.

Endor (Israël) fait de la « social physics », un autre nom pour la psychohistoire, cette pseudo-science inventée par Isaac Asimov, qui permet de prévoir à partir de signaux faibles ce que nous achèterons le mois prochain. Enfin, peut-être !

Cognitive Matchbox (2016, France) améliore l'aiguillage des clients vers les bons agents dans les centres d'appels en fonction des émotions et des besoins. Cela rappelle ce que fait Batvoice. Ca tombe bien puisque les clients aiment bien tomber sur de vraies personnes lorsqu'ils appellent une société et pas sur un robot qui va leur répéter ce qu'ils ont déjà vu sur leur site web !

Segmentify (2015, Turquie, 1M€) fait du marketing en ligne par email et application mobile pour harceler les clients et leur "*proposer les produits qu'ils aiment*". Evidemment, cela s'appuie sur du machine learning. Bref, c'est surtout un outil de recommandation.

[609] Nombreux details ici : http://www.businessinsider.fr/us/mercedes-electric-vision-van-drone-delivery-service-photos-2017-3/.

Skymind (2014, USA, $6,3M) a été créée par des anciens de Vicarious. Elle propose une solution open source en Java – Deeplearning4j.org – capable d'analyser des flux de données. Elle est notamment utilisée dans la détection de fraude, le commerce et le CRM.

Logistique

La logistique du commerce et plus généralement la « supply chain[610] » peuvent exploiter à de nombreux endroits des briques d'IA :

- La **détection des tendances** dans les réseaux sociaux qui permettent de prédire des ventes soudaines et l'engorgement de chaines logistiques comme les toupies à main en 2017.
- Dans l'automatisation des **processus de dédouannement** avec des solutions juridiques développées avec **IBM Watson** et du traitement du langage.
- Le suivi et la surveillance des **conteneurs maritimes** et des wagons dans le rail avec **Traxens** (2012, France, 1,5M€).
- L'optimisation des flux dans les **transits aériens** qui représentent 35% de la valeur du commerce mondial pour 1% du volume.
- L'optimisation de trajets de livraisons proposée notamment par **LogiNext** (2014, USA, $10,6M). L'anglais **Deliveroo** utilise sa solution Frank qui est censée avoir réduit les délais de livraison de 20% au Royaume Uni à partir de 10 000 restaurants et pour 15 000 livreurs.
- La robotique de tri et de dispatch de colis dans les entrepots de livraison, comme avec **Exotec Solutions** (2015, France, $21,2M) est ses robots de transports de produits associés à des postes de préparations de commandes[611].

- L'**inspection visuelle d'entrepots** comme avec **Qopius** (2015, France, 1,8M€).
- L'**inspection de linéaires** dans le retail avec **Bossa Nova** (2005, USA, $69,5M) avec son robot qui aligne des kilomètres dans les rayons des magasins pour analyser la présence des produits dans les linéaire et permettre le réassortiment rapide. Par contre, il ne remplit pas encore les rayons. Chaque chose en son temps ! Il serait en test dans une cinquantaine de magasins Wallmart depuis fin 2017 (vidéo) et n'aurait pas encore renversé d'enfant en bas âge ce qui est plutôt bon signe. Dans sa prochaine version, il surveillera les clients. Ou pas.

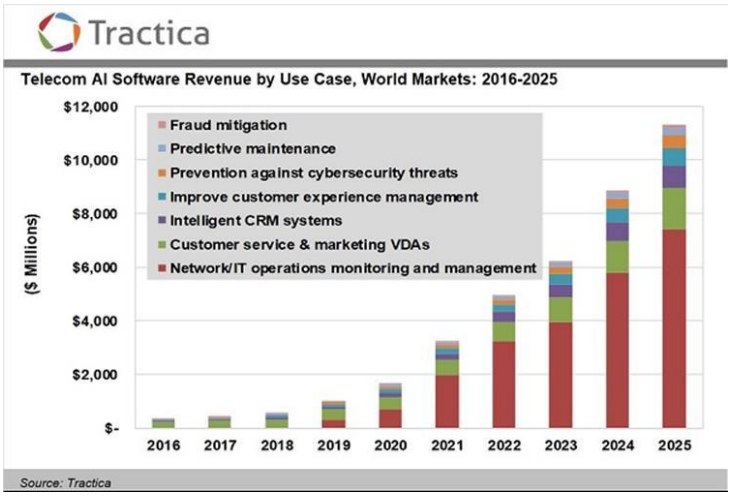
[610] Voir 6 Applications of Artificial Intelligence for your Supply Chain., octobre 2017.

[611] Ces exemples sont en partie extraits de Artificial Intelligence in Logistics, un rapport commun de DHL et IBM, 2018 (45 pages).

Télécoms

Le marché des télécoms est l'un des premiers à être concerné par l'usage de l'IA dans son exploitation comme dans ses offres et leur marketing. Et pour cause, les opérateurs ont à gérer des infrastructures complexes et distribuées. Ils génèrent aussi de gros volumes de données avec l'activité des réseaux et de leurs utilisateurs.

L'offre d'IA y est cependant moins lisible que sur les autres marchés. Elle est concentrée sur quelques acteurs à commencer par les grands équipementiers (**Huawei**, **Cisco**, **Nokia**, **Ericsson**, **ZTE**), puis diverses startups et des prestataires de services et enfin, les opérateurs télécoms eux-mêmes, qui développent souvent leurs propres solutions sur mesure ou conjointement avec leurs équipementiers et prestataires de services[612]. Ils jouent souvent le rôle d'intégrateur. D'autres comme **Free** ont tendance à exploiter des logiciels open source et à développer les leurs.



Parmi les éditeurs de solutions logicielles à base d'IA destinées aux opérateurs télécoms, domineront essentiellement ceux qui fournissent des outils de gestion des opérations réseau et IT, suivis par ceux qui proposent des outils liés à la relation clients[613].

Comme partout, l'IA est une grosse boîte à outils qui est exploitée, application par application, et de manière disparate. Il n'y a pas encore de gosplan ou de méta-architecture d'IA pour gérer l'ensemble des systèmes d'information des opérateurs télécoms.

Les principaux éditeurs de logiciels des télécoms sont soit génériques soit spécifiques à ce marché, avec **Afiniti** (association agents/clients en centres d'appels), **AIBrain** (IA générique de raisonnement, moteurs de règles et réseaux d'agents), **Anodot** (détection d'anomalies en machine learning), **Aria Networks** (optimisation de réseaux), **B. Yond** (et sa plateforme d'analytics et de supervision de réseaux Infinity), **Cardinality** (analytics engines), **Guavus** (analytics temps réel), **Intent HQ**, **Ipsoft** (chatbot d'assistance d'agents de helpdesk), **Skymind** (déploiement distribué de solutions d'IA), **Subtonomy** (monitoring de réseau mobile), **Tupl** (monitoring de réseaux), **Sysmech** (outils de gestion de réseaux) et **Wise Athena** (outil de pricing).

Réseau et exploitation

Le déclencheur d'un usage intensif de l'IA chez les opérateurs sera probablement le déploiement et l'exploitation de réseaux sans fil 5G dans les douze ans qui viennent. Le niveau de complexité de ces réseaux augmente d'un ordre de grandeur par rapport à la 4G.

La 5G va voir se généraliser des concepts tels que le **SDN** (Software Defined Network), les **SON** (Self-Organized Networks), et la **NFV** (Network Function Virtualization) [614].

---

[612] Comme Nokia qui est partenaire de China Telecom dans l'IA. Voir Nokia and China Mobile to set up joint AI*5G lab for further research using artificial intelligence and machine learning in 5G networks, juillet 2018.

[613] Source : étude marché de **Tractica**, 2018, vue dans Report says AI Sales in Telecom to Reach $36,7B By 2025, avril 2018.

[614] Et cela comprend aussi **ONO** (Online Network Optimization) qui est proposé dans Model-Driven Artificial Intelligence for Online Network Optimization 2018, (10 pages) et qui consiste à configurer dynamiquement les réseaux avec du deep learning alimenté par les données des SDN.

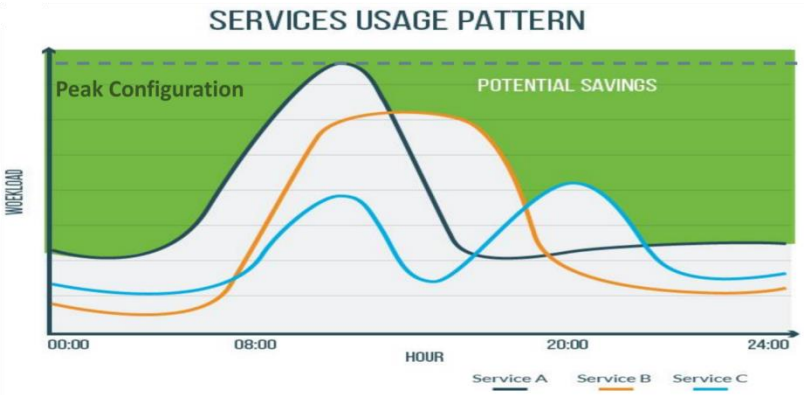
Tout cela pour décrire des architectures de pilotage de réseaux reconfigurables qui reposent sur des serveurs génériques et sur une gestion très dynamique des réseaux et de leurs infrastructures. Les réseaux vont pouvoir se reconfigurer dynamiquement en fonction du trafic[615].

Les réseaux 5G vont nécessiter de prédire l'évolution du trafic avec une forte granularité pour allouer les fréquences dans la 5G notamment dans les zones très denses, et pour maximiser l'usage de nombreuses fréquences traditionnelles (en-dessous de 5 GHz), pré-millimétriques (5 à 28 GHz) et millimétriques (28 GHz et au-delà).

D'autres enjeux sont à prendre en compte comme la réduction de la consommation d'énergie des serveurs hors des heures de pointe d'usage du réseau. Tout cela en temps réel, au gré de l'évolution de la demande.

Les réseaux vont aussi voir augmenter la part des vidéos qu'ils transportent, avec des exigences non négligeables en termes de qualité de service qui va nécessiter encore plus de coordination entre les techniques d'adaptive streaming et les liens avec les CDN (Akamai and co).

L'IA peut intervenir pour prédire les évolutions de la demande de télécommunications et aider les opérateurs à préparer les infrastructures à s'y adapter, de plus en plus autonomatiquement. Cela nécessitera d'ailleurs probablement des techniques d'apprentissage par renforcement pour affiner les modèles prédictifs au fil de l'eau. C'est une approche retenue par **China Telecom**[616] (*ci-contre*).



En amont du pilotage, l'IA – et même un jour, le calcul quantique – peut intervenir pour optimiser le placement des antennes et minimiser leur nombre tout en maximisant leur couverture, le problème mathématique sous-jacent étant celui de la résolution du théorème des quatre couleurs.

La supervision des réseaux des opérateurs va utiliser de plus en plus de machine learning. Le Corée **SK Telecom** utilise d'ailleurs depuis 2017 une telle solution pour gérer ses opérations réseaux et améliorer la qualité de service et l'état de ses infrastructures, dénommée Tango (vidéo).

Enfin la maintenance des infrastructures peut aussi faire appel à du machine learning pour identifier les infrastructures à maintenir ou remplacer avant que des pannes surviennent. **Sysmech** (1994, UK) propose ainsi des outils de supervision de réseau à base de machine learning qui peuvent détecter les émetteurs du réseau qui sont en panne pour les redémarrer.

Cela passe même par l'inspection visuelle des antennes par des drones comme chez **Aerialtronics** (2012, Pays-Bas, 3M€), une technique qui est notamment utilisée par **AT&T** aux USA (vidéo) depuis 2016.

Offres marchés

Le second domaine où l'IA intervient chez les opérateurs télécoms est dans leurs offres commerciales, surtout grand public.

---

[615] Voir aussi Defining closed-loop AI mechanisms for network management de l'ETSI, 2018 (55 slides) ainsi que Artificial Intelligence for 5G: Challenges and Opportunies de Merouane Debbah du centre de recherche de Huawei en France, 2018 (64 slides).

[616] AI and other emerging ICT technologies bring new development opportunities for telecom operators de Xiaoou Liu, China Telecom, avril 2018 (32 slides).

Du côté du fixe, les outils de **recommandation de vidéos** s'appuient souvent sur du machine learning sont légion, même s'ils sont souvent issus d'éditeurs de logiciels indépendants. Les box sont aussi commandables à la voix. Par contre, les opérateurs interviennent moins sur les mobiles eux-mêmes du fait de la maîtrise des outils logiciels par les plateformes Android et iOS.

Orange propose son agent vocal **Jango** qui est utilisé dans différents objets connectés ainsi que dans leur offre Orange Bank.

Au-delà de la recommandation de contenus divers, l'IA peut intervenir à plusieurs endroits de la chaine produit, que ce soit pour de l'upsell ou du crosssell, dans les boutiques sur le terrain et dans les applications de web-to-store.

La situation devrait changer avec la 5G qui va voir les opérateurs télécoms ayant pignon sur rue dans les entreprises. Ils vont mettre un nez dans des marchés verticaux qui seront très consommateurs de briques logicielles d'IA : avec les véhicules connectés et autonomes, dans la santé et la ville intelligente. Les opérateurs sont impliqués dans ces domaines comme des fournisseurs d'infrastructures. Mais parfois, ils s'engagent avec du service et de l'intégration comme chez Orange Business Services ou chez Bouygues Telecom via leur filiale Objenious.

Les opérateurs mettent aussi en place des solutions à base d'IA pour la **détection de différentes formes de fraude**, surtout aux USA. Cela concerne l'usurpation d'identité avec de fausses pièces d'identités utilisées pour souscrire des abonnements.

Chez **Deutsche Telekom**, on utilise depuis 2018 les logiciels biométriques de l'Américain **Nuance** pour réaliser cette authentification des utilisateurs lorsqu'ils appellent le support en ligne. La solution permet aussi de faire sa demande en langage naturel pour être orienté vers la bonne personne, au lieu de passer par une interminable suite de questions et de numéros à saisir. Cette solution de Nuance est également déployée en Turquie par **Vodafone**.

Relation client

On pense évidemment en premier aux chatbots qui sont maintenant très courants chez les opérateurs télécoms. On en trouve chez **Bouygues Telecom**, chez **Orange**, chez **Telefonica** et **Vodagone** (avec Tobi[617]). Ils permettent notamment d'accéder aux bases de connaissance du support technique.

Aux USA, on en trouve chez **Spectrum**, l'opérateur issu de la fusion de Charter Communications et Time Warner Cable, propose l'assistant virtuel Ask qui gère le support commercial et technique. Le câblo-opérateur **Century** utilise depuis 2016 son agent Angie développé avec **Conversica** (2007, USA, $72M) et qui analyse 30 000 mails entrants par mois.

Tous ces outils couplés aux indicateurs d'utilisation des services télécoms fixes et/ou mobiles permettent de faire du machine learning pour identifier les clients succeptibles de quitter le service (le « churn »). A charge ensuite de 1) comprendre la situation exacte du client et 2) lui proposer une éventuelle offre adaptée différente du package dont il dispose.

L'autre approche consiste à associe chatbots et agents humains dans les centres d'appels, ce que propose **Afiniti** (2006, USA, $137M).

[617] Voir http://labs.vodafone.co.uk/case-studies/tobi et leur vidéo.

En dernier lieu, citons l'usage des robots **Pepper** pour l'accueil dans les boutiques de Softbank au Japon, qui y sont déployés depuis 2014 (*ci-contre*, dans celle d'Otomesando à Tokyo en octobre 2014). Quel retour en avons-nous ? Le déploiement était explicable par le fait que l'équipe conceptrice de Pepper issue de l'acquisition du Français Aldebaran Robotics fait aussi partie du groupe Softbank. Mais les scénarios d'usage de Pepper en boutique restent limités. Qu'en reste-t-uil une fois passé l'effet de surprise et de découverte initiale ?

Médias et contenus

Les médias font partie de ces métiers qui ont été particulièrement bousculés par l'irruption des outils numériques, d'Internet et des médias sociaux. Leur chiffre d'affaire a baissé, leurs revenus publicitaires ont en parti migré vers d'autres acteurs, que ce soit les GAFA, les services en ligne d'offres d'emplois ou dans l'immobilier.

Les moyens baissant, ceux qui sont alloués aux journalistes pour mener leurs enquêtes ont fondu d'autant. Nombre de médias ont décliné, surtout dans la presse quotidienne nationale et régionale.

L'adoption de nouvelles méthodes de travail ne s'est pas faite sans mal. Les rédactions digitales étant trop souvent séparées des rédactions historiques. Les premières ont adopté des méthodes peu recommandables, republiant des informations sans prendre le temps d'enquêter, générant des effets de caisse de résonnance à ce que l'on appelle maintenant les vraies fausses nouvelles. Les seconds ont de leur côté ignoré les outils permettant à leurs écrits d'être mieux diffusés.

Sur ce arrive la vague de l'IA qui entraîne tout sur son passage et qui peut à son tour bouleverser une fois de plus les médias. Avec la crainte qu'elle génère une nouvelle vague déflationniste du côté de l'emploi.

Nous allons donc voir ici comment les médias, et surtout la presse, peuvent tirer parti de l'IA à la fois pour la production de contenus, pour leur diffusion et pour leur monétisation[618] et si cela prête à conséquence. L'IA est comme Internet et Google Search. Elle peut entraîner le meilleur comme le pire. Elle permet aussi bien d'améliorer la qualité de ses contenus que de se désengager de ce point de vue-là.

L'IA peut notamment aider les journalistes à analyser les données et détecter des tendances à partir de sources d'informations multiples allant des sources ouvertes habituelles aux sources inédites comme les données publiées par Wikileaks. Elle peut aussi aider à convertir les données en texte, les textes en contenus audio et vidéo. Elle permet d'analyser des objets, des images, de reconnaître des personnes dans des photos, ce qui est très utile pour comprendre des situations captées par des photo-journalistes.

Nous évoquerons aussi le rôle de l'IA dans d'autres types de contenus, en premier lieu la musique et les jeux vidéo. L'IA apporte de nouveaux outils de génération de contenus qui font peur à certain. En fait, il s'agit d'une extension de la palette des outils mis à la disposition des créatifs et à charge pour eux de s'en emparer[619].

---

[618] Voir Et si les médias redevenaient intelligents ?, mai 2017.

[619] L'intelligence artificielle créative…vers de nouveaux horizons d'Olivier Reynaud, septembre 2018.

Recherche

Les nombreux moteurs de recherche documentaires génériques du marché peuvent être exploités par les journalistes s'ils peuvent s'en équiper. Il en existe quelques-uns qui sont dédiés aux journalistes, comme Salient de **Lore.ai** (2016, France/USA), un outil d'analyse de contenus qui permet de réaliser des recherches, des liens entre documents, de les classifier et de les exploiter, le tout sur plusieurs sources et plusieurs langues.

Newsbridge (2016, France) réalise de l'indexation de vidéos par analyse des voix qu'elles contiennent avec du speech-to-text. Cela permet ensuite de faire facilement des recherches dans ses rushes de reportages ou dans les vidéos déjà publiées.

Ecrit

Nombre de startups sont apparues qui automatisent la production de contenus. Comme presque partout dans l'IA, ce qu'elle produit ne vient pas de nulle part mais réutilise des contenus existants créés par de vrais gens. Et elle les assemble en observant la manière dont les contenus existants sont assemblés.

Nous avons pu balayer quelques startups spécialisées dans la génération automatique de textes et de résumés dans la rubrique sur le traitement du langage. Une partie d'entre elles ciblent les médias, pour ce que l'on appelle le « robot journalism » comme le **Syllabs** (2006, France, 2M€) et sa solution Syllabs Media utilisée chez Radio France, à l'AFP ou dans le journal Le Monde, ou encore **Textomatic** (2010, Allemagne, $40M). Syllabs Media est voisin de la solution d'Yseop et génère des textes à partir de données structurées et chiffrées. Le journaliste doit ensuite modifier le texte pour lui ajouter de l'émotion et du vivant.

Très souvent, les startups qui ciblent intialement le marché des médias s'en détournent progressivement ou partiellement parce que ce marché est moins solvable et dynamique que ceux de la finance ou du marketing.

Les robots rédacteurs ne font pas de véritable journalisme. Ils génèrent des textes répétitifs sur de gros volumes de données comme pour produire les résultats d'élections à l'échelle locale, dans le cadre de compétitions sportives ou pour la météo.

Ils transforment le plus souvent des données numériques en phrases avec des templates plus ou moins flexibles. Mais les générateurs de langage peuvent de plus en plus tenir compte du contexte des données pour choisir les bonnes formulations.

C'est le cas avec la publication en mai 2017 d'un article du Los Angeles Times annonçant un tremblement de terre d'échelle 3.8 (vidéo)[620]. L'article n'a pas été écrit par un système à base d'IA mais par un petit programme dénommé ClickBot développé par un journaliste du LA Times. Un journaliste a ensuite complété l'article à la mano.

[620] L'article en question : http://www.latimes.com/local/lanow/la-me-earthquakesa-earthquake-39-quake-strikes-near-view-park-windsor-hills-calif-onvisi-story.html.

L'agence **Associated Press** publie depuis 2015 des dépêches créées par des robots journalistes pour les annonces standardisées, notamment dans l'actualité financière (vidéo).

Le traitement de gros volumes de données générés par des sources telles que Wikileaks nécessite aussi des outils spécifiques. Ils sont souvent développés à bas cout à partir des nombreuses briques logicielles en open source du marché. Les rédactions des médias ne sont pas suffisamment fortunées pour se payer les services d'une grande ESN ou d'un SAP !

La simple digestion de vidéos est trop longue pour une rédaction dans le print. Les outils de transcript de vidéos en texte sont donc les bienvenus. C'est une fonction standard dans **YouTube** (*ci-dessous*) !

D'autres solutions de génération de contenus dédiés aux médias ont vu le jour sur d'autres types de contenus. **Valossa** (2015, Finlande, $2,7M) propose ainsi une solution en cloud de reconnaissance d'images dans les vidéos adapté aux besoins des broadcasters.

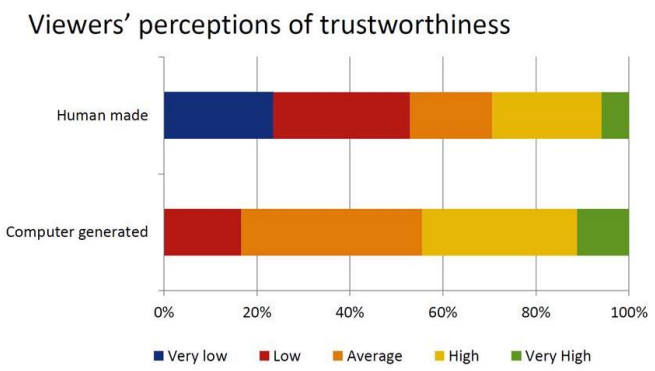
Elle permet l'interprétation de vidéos, détecte les personnes, leur verbatim et les thèmes couverts (*ci-dessous à droite*). Elle ajoute des métadonnées aux scènes analysées exploitables dans les outils d'analytics voire pour les générateurs de guides de programmes.

Lobster (2013, UK, $2M) fait de la curation de contenus issus des médias sociaux pour alimenter les agences de communication et entreprises

Le breton **Mediawen** (2014, France) gère la traduction de contenus vidéo en temps réel en s'appuyant sur IBM Watson puis text to speech, en voix de synthèse ou sous titrage.

Des chercheurs de **Thomson Reuters** et **Alibaba** génèrent de leur côté des news inédites (« scoops ») à partir des flux détectés dans Twitter[621] dans l'application Reuters Tracer. Le système repère les nouvelles, les classifie, les annote, les mets dans l'ordre, rédige un article et le publie. Et il le fait plus vite que les agences de presse traditionnelle. Et même s'il faut peut-être éditer le résultat, le processus sera toujours plus rapide.

Une étude américaine montre que les lecteurs font plus confiance à des articles produits par des robots que par des journalistes[622]. Cela mérite évidemment un peu de recul car les articles produits par les robots journalistes ne font le plus souvent que transformer des données chiffrées en phrases et ne véhiculent donc pas d'opinion ou de jugement de valeur. Sauf … si les données sont fausses ! Et, sans surprise, les journalistes n'aiment pas les robots journalistes[623] !



Knowhere (2015, USA, $2M) est une startup qui résout le problème en sélectionnant des articles structurellement neutres. C'est une sorte d'outil de curation qui jauge de la véracité des articles. Il sélectionne d'abord les sujets en fonction de leur popularité, ce qui est déjà un biais humain soit dit en passant, ne retient que les sujets cités dans au moins cinq sources fiables, en éliminant les sites conspirationnistes. Un article synthétique est ensuite automatiquement réalisé par mash-up des articles sélectionnés en retenant les faits et en éliminant les avis. Un journaliste intervient en bout de course pour valider l'ensemble. Notez cependant que la source de tout cela sont des articles écrits par des journalistes en chair et en os.

L'équipe du Medialab de Denis Teyssou à l'**AFP** a mené plusieurs projets qui aident les journalistes de l'agence à détecter les fake news en recherchant notamment l'origine exacte de photos et vidéos qui peuvent ne pas refléter les événements qu'ils sont censés décrire[624].

Ces initiatives sont complétées par **DeepNews.ai**, un projet lancé par Frédéric Filloux, visiting professor et chercheur à Stanford depuis mi 2017. Celui-ci opère sur un mode différent en sélectionnant sur Internet les articles les plus pertinents sur les sujets d'actualité. L'algorithme s'appuie sur l'analyse de la profondeur du traitement du sujet, sur l'expertise mise en avant, les qualités de l'analyse et les moyens mis en œuvre. La dernière mouture de l'outil s'appuie sur un réseau convolutionnel. Il est surtout destiné à des plateformes d'agrégation[625].

---

[621] Voir Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data, novembre 2017 (11 pages).

[622] Source du schéma à droite : The Artefacts of Automated Journalism: Producers' Perspectives and Audience Assessments de Neil Thurman.

[623] Voir Human journalists hate robot journalists, says new report, mars 2017, qui propose au passage un petit test de détection de brèves rédigées par des robots et des journalistes. Il n'est pas trop difficile d'obtenir 5 sur 5 au test. Ce qui est rassurant pour les journalistes !

[624] Voir AFP Medialab: information verification and user experience at the core of newsroom innovation et l'intervention de Denis Teyssou dans une table ronde au Web2day de juin 2018 à partir de la seizième minute.

[625] Voir A progress report on Deepnews.ai, mars 2018 et Deepnews.ai, Progress Report #2: It works, octobre 2018.

Enfin, citons ces IA qui génèrent des nouvelles ou des poèmes[626]. Nombre de commentateurs s'en émerveillent alors qu'il ne s'agit que de systèmes probabilistes et combinatoires qui n'ont aucune intelligence symbolique ni capacité émotionnelle.

Le fruit de la combinatoire et d'une partie de hasard suffit à générer des créations et à faire illusion. On oublie évidemment que l'Homme fait le tri en sortie de machine pour ne conserver que cela qui a l'air exploitable et publiable. Les contenus poubellisés sont rarement présentés dans ces productions !

Musique

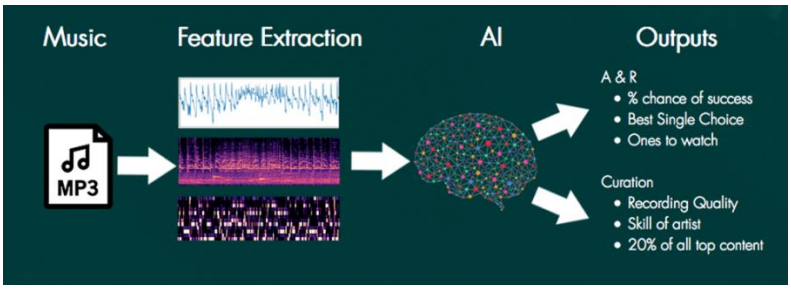
L'industrie musicale tire aussi parti de l'IA pour la production et la diffusion de contenus. C'esr une grande consommatrice de réseaux génératifs. Les outils de création de musique artificielle entraînés par l'ingestion de wagons de répertoires musicaux sont légion. Pour l'instant, ils ont bien du mal à créer de la musique de qualité.

C'en est encore au stade de la musique d'ascenseur ou de jingles publicitaires. Mais cela progresse doucement. L'ADN du hit n'a pas encore été trouvé et laisse encore de la place aux véritables artistes.

- **Landr** (2012, Canada, $9M) propose un service en cloud d'automatisation du mixage audio, qui va créer des morceaux de musique agréables à l'écoute (vidéo).
- **Popgun.ai** (2017, Australie) utilise le deep learning pour apprendre les bonnes règles musicales à partir de compositions humaines et pour enrichir des compositions existantes. La démonstration de leur prototype Alice est sympathique (vidéo) mais pas forcément facile à mettre en pratique.
- **Amper Music** (2014, USA, $9,1M) est un site en ligne, pour l'instant gratuit, qui compose automatiquement de la musique via de l'IA (vidéo). Il faut tout de même le paramétrer pour indiquer ses souhaits en termes d'instruments, de tempo, de style et de durée. C'est pratique pour créer des compositions qui vont alimenter les vidéos de démonstration de startups, et éviter les habituelles musiques d'ukulélé qui les accompagnent régulièrement. La startup a plusieurs concurrents directs comme **Jukedeck** (UK, $3,4M, exemple), **Melomics Media** (2012, USA, exemple), **Aiva Technologies** (2016, Luxembourg, $768K) (exemple 1 de musique symphonique et exemple 2 de musique rock), **Melodrive** (2017, Allemagne, exemple).
- **Pacemaker** (2011, Suède, $4,5M) est un DJ à base d'AI qui exploite les contenus de services de musique en ligne comme Spotify.
- **MXX** (2015, UK) propose une solution logicielle qui édite une musique automatiquement pour la synchroniser avec une vidéo faisant gagner beaucoup de temps au monteur.
- **Muzeek** (2015, USA) génère directement une musique originale et la synchronise avec une vidéo. Le marché visé est surtout celui de la création de spots publicitaires.

[626] Voir par exemple When an AI Goes Full Jack Kerouac - A computer has written a "novel" narrating its own cross-country road trip, de Brian Merchant dans The Atlantic, octobre 2018. Et puis aussi Turing Test Passed : Using Computer Generated Poetry, janvier 2015. L'auteur ne sait visiblement pas ce qu'est le test de Turing ! C'est un test portant sur un dialogue avec une machine qui serait impossible à distinguer d'un homme. Ici, il ne s'agit pas de dialogue. Voir aussi le générateur de poèmes de Ray Kurzweil : Ray Kurzweil's Cybernetic Poet: How it works.

- **Musiio** (2018, Singapour) a développé un outil qui permet aux labels d'édition d'artistes de sélectionner les artistes, les morceaux à éditer et les singles à mettre en avant. Il s'entraîne avec l'ingestion de dizaines de milliers de musiques avec du deep learning.



  Cela rappelle le projet du chercheur de la BBC Armand Leroi publié en 2017 et ayant donné lieu à un documentaire de la chaîne[627]. Le projet était "inconclusif", à savoir que si le machine learning avait bien permis de déterminer quelques caractéristiques des morceaux qui plaisaient selon les époques, il n'avait pas permis de créer pour autant un nouveau hit.

- **Nami Lab** (2016, Italy, 250K€) a développé l'application Yalp qui fait de la transcription en partition commentée de toute musique. Cela peut aider les musiciens, notamment les guitaristes, à reprendre des morceaux de répertoire.

- **Google** a créé son propre générateur de sons à base de deep learning, Nsynth (Neuron Synthetizer), issu du projet Magenta. Il est couplé à un instrument de pilotage, Nsynth Super (vidéo) qui pourrait un jour faire partie de la panoplie des compositeurs et DJs. Mais ce n'est pas encore un produit. C'est juste un projet de chercheurs.

- Le laboratoire **Sony CSL** situé à Paris a créé Flow Machines, une IA génératrice de pop-music. En 2016, elle créait Daddy's Car, un succédané des Beatles sauce Georges Martin s'étant emmêlé les paluches dans la table de mixage d'Abbey Road ou Phil Spector ayant mélangé un peu trop de drogues douces et la bière de Brett Kananaugh (vidéo)[628]. Plus récemment, elle composait l'album « Hello World » (vidéo) mais avec l'aide d'artistes.

- Le chanteur de The Voice en Norvège **Thomas Holm** créait une chanson de Noël via une IA exploitant une bibliothèque de chansons du même type, avec paroles et musique[629].

- Enfin, on peut apprécier le projet open source français **AdBlock Radio** qui sert à éviter la publicité dans l'écoute de la radio. Il a été lancé par Alexandre Storelli, un ingénieur de l'X. Il n'est évidemment pas certain que ce logiciel ait un modèle économique.

Photo

La photo fait l'objet d'innombrables solutions exploitant l'IA, à commencer par les applications photos des smartphones qui savent gérer la mise au point en analysant les scènes photographiées. Elle sert aussi au classement de photos et à la détection de leur qualité. Les outils de gestion de bases de photos intègrent progressivement des briques d'IA pour les rendre un peu plus smart.

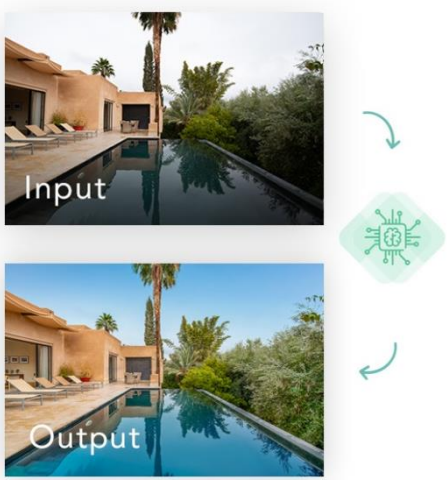
---

[627] Voir The Secret Science of Pop, 2017.

[628] L'histoire évite d'indiquer combien de morceaux créés par l'IA en question ont du être poubellisés avant la production de celui qui a été retenu ! In fine, l'Homme conserve ce qui lui semble être pertinent. L'IA n'a pas encore cette capacité de jugement artistique et émotionnelle. Ces IA créatrices doivent être surtout percues comme de nouvelles palettes pour les compositeurs. Les IA créatrices fonctionnent encore par apprentissage par renforcement, ce dernier provenant des créateurs humains.

[629] Voir Artificial Intelligence Meets Christmas Spirit, 2017.

Adobe ajoute régulièrement des fonctionnalités à base d'IA dans ses logiciels comme Lightroom, qui font de plus en plus appel à des ressources de calcul dans le cloud. Mais ils ne sont pas seuls sur ce créneau[630].



Il faut aussi compter avec **Meero** (2016, France, $63,4M) qui est spécialisé dans l'usage d'IA pour le traitement et la retouche automatique de photos et qui vise les marchés professionnels comme l'immobilier, le e-commerce et la restauration. Ce n'est pas anodin : cette startup qui est parmi les mieux financées dans l'IA en France. Elle gère la production de photos en s'appuyant sur des photographes externes mais toute la post-production est automatisée, ce qui permet de réduire le coût des prestations. La startup revendique avoir déjà 40 000 clients avec une empreinte mondiale.

Banuba (2016, Hong Kong, $5M) propose un SDK aux développeurs d'applications mobiles pour interpréter le contenu d'images ou les modifier. C'est une belle boîte à outils multifonctions qui vise surtout le marché des développeurs d'applications de réalité virtuelle ou augmentée.

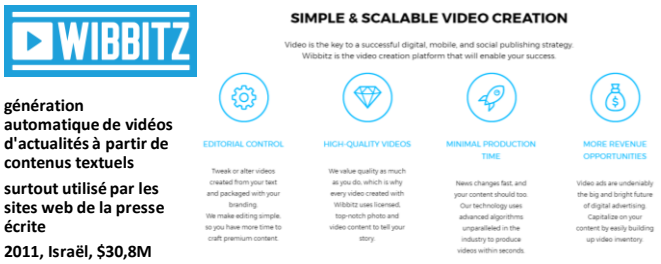
Elle sert à détecter les émotions, suivre le regard, créer des modèles 3D de visages intégrables dans des applications de réalité virtuelle, faire du mocap (motion capture, capture de mouvements, vidéo), détecter et modifier la couleur de la peau et des cheveux, améliorer les visages et enfin, enlever le fond derrière une personne.

Vidéo

Passons à la génération de vidéos à base d'IA où la créativité est aussi grande que dans celle de la musique. La plupart de ces outils font du montage automatique de vidéos et photos existantes. Les grands éditeurs du marché comme **Adobe**, **Avid** et **Elemental** (filiale d'Amazon) ajoutent fonctions après fonctions de traitement automatique des vidéos pour faire gagner du temps aux monteurs[631].

Il vous est peut-être déjà arrivé de tomber sur des vidéos sur YouTube se présentant sous la forme de slideshow avec une voix off robotisée lisant un texte. Ce sont des bots de génération de spam vidéo ! Bref, le pire [632]! Dans d'autres cas, ce sont des vidéos qui se lancent toutes seules dans des journaux en ligne et qui lisent les articles. C'est bien pour les mal voyants mais pénible pour les autres ! Heureusement, la génération automatique de vidéos a d'autres applications que voici.

Wibbitz (2011, Israël, $30,8M) est connu pour sa solution qui génère automatiquement des vidéos d'actualité à partir de contenus textuels et de scrapping de contenus vidéos. Elle est surtout utilisée par des médias de la presse écrite qui ont besoin de compléter leur production par des contenus vidéo dans leur version en ligne[633]. Elle a comme concurrent direct **Wochit** (2012, Israël, $28,8M).



---

[630] Voir Lightroom CC 2.0: What's new, and where is it headed? par Jeff Carlson, octobre 2018. Dans cette version, on retrouve la détection de visages permettant de trier les photos par personne. On peut aussi faire des recherches dans ses photos en en décrivant le contenu (ciel, mer, bâteau, avion, etc).

[631] Voir Top Video Artificial Intelligence and Machine Learning at NAB 2018, mai 2018.

[632] Voici un exemple de ce genre de vidéo : https://www.youtube.com/watch?v=Kkwq1Uht_A0.

[633] C'est ce que fait également **Oovvuu** (2014, Australie, $3,8M) .

Magisto (2009, Israël, $22,5M) a créé un outil de montage de vidéos dans lequel on télécharge ses éléments photos et vidéos existants, qui les analyse et réalise ensuite un montage automatique à partir d'un choix préalable de style et de musique d'accompagnement. Ce sont peut-être eux qui proposent ces satanées musiques de yuku-lele aux startups IOT ! L'algorithme analyse la dimension visuelle, audio et narrative de la vidéo.

Cela cible surtout la création de vidéos à des fins de marketing (vidéo). C'est assez impressionnant et ça marche en mode web et mobile.

Graava (2014, USA, $2,3M) a créé une caméra qui fait elle-même son propre montage automatique. Et dans un registre voisin, **Soloshot** (2011, USA) a conçu une caméra robotisée pour capter automatiquement les objets en mouvement, ce qui n'a rien d'extraordinaire mais peut éventuellement s'appuyer sur de la reconnaissance d'image à base de réseau convolutionnel.

Arraiy (2016, USA, $13,9M) génère des effets spéciaux automatiquement pour le cinéma, la TV et les jeux vidéo. Le principe consiste surtout à extraire par détourage des personnages du fond de l'image pour les incruster dans d'autres scènes, sans passer par l'usage d'un fond vert ou bleu. Cela reprend le très vieux procédé du rotoscope. Le tout utilise des réseaux de neurones entraînés avec de gros volumes d'archives de vidéos. Le projet est en cours de développement.

Triller (2015, USA, $11,9M) a publié une application mobile qui peut créer une vidéo montée exploitant d'un côté de la musique de répertoire et des vidéos filmées par les utilisateurs. Il va synchroniser le tempo du montage avec celui de la musique. C'est déjà un phénomène chez les jeunes youtubers (vidéo). Mais c'est quelque peu déceptif et ce n'est pas qu'une question d'IA.

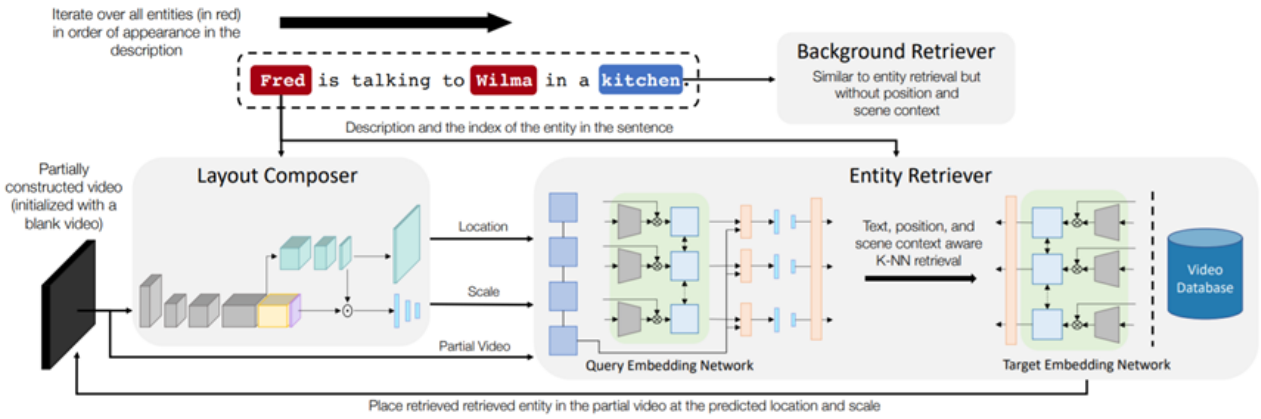
IBM Watson a été utilisé pour créer un système de génération automatique de bande annonce de films, testé notamment pour un film inconnu, Morgan (vidéo).

ScriptBook (2015, Belgique, $1,4M) ne génère pas de vidéo analyse les scripts de films automatiquement pour aider à les choisir avec son logiciel Script2Screen. Elle prévoit même le succès des films au box office, ce qui est quelque peu présomptueux sans connaître le réalisateur et les acteurs.

Le décalage entre un script et le produit fini d'un film est quand même généralement significatif. Malgré tout, leur premier test ne donne aucun faux positif sur un échantillon de 65 films d'une maison de production d'Hollywood et élimine plus de scripts ayant généré des échecs que les humains impliqués dans le processus.

Le projet **Imagine This!** des Universités de l'Illinois et de Washington (Seattle) génère automatiquement des dessins animés des Flintstones à partir du script[634]. On est dans le registre des réseaux génératifs, c'est assez sophistiqué mais le résultat n'est pas ébouriffant pour autant car le mode d'animation généré est très simpliste dans ce type de dessin animé (vidéo).

[634] Voir Imagine This! Scripts to Compositions to Videos, avril 2018 (22 pages).



Enfin, le plugin **Trint** d'Adobe Premiere est capable de générer automatiquement le sous-titrage des contenus vidéo via sa fonction de speech-to-text[635]. Après, il ne reste plus qu'à corriger les erreurs qui ne manquent pas d'être générées dans ce genre d'outil, mais cela fait tout de même gagner du temps

Jeux vidéo

L'IA est aussi utilisée dans les jeux vidéo et, en fait, depuis pas mal de temps. Nombre de jeux vidéos reproduisent des mondes complexes avec un grand nombre de personnages en s'appuyant sur des systèmes multi-agents qui animent des personnages respectant leur propre mode de fonctionnement et interagissant entre eux. Des personnages de jeux à part entière peuvent être aussi animés par des IA comme l'a expérimenté **OpenAI** dans le jeu **DOTA 2** (voir la partie).

D'autres jeux comme OverWatch de **Visor** conseillent le joueur pour améliorer sa pratique de jeu en l'observant, et avec des indications en temps réel.

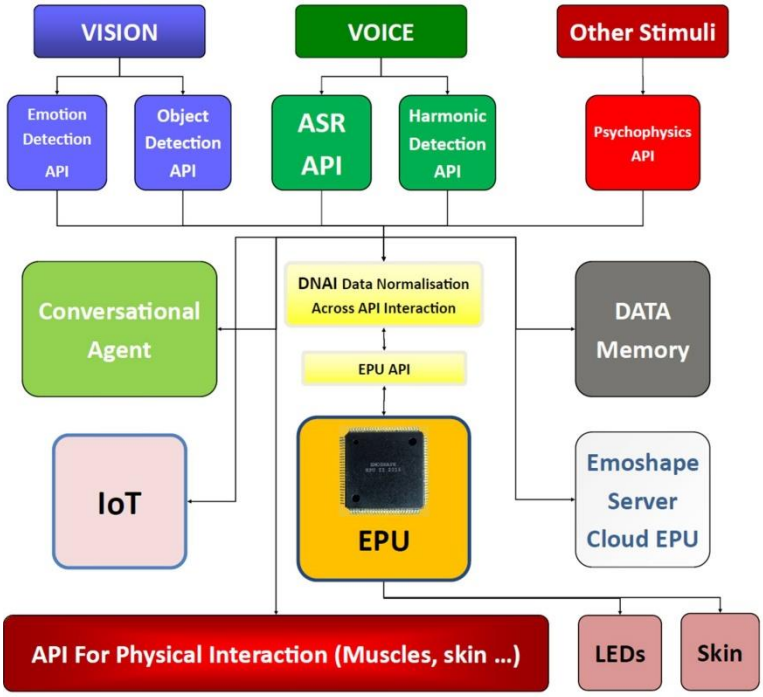
Mais reste à rendre tout cela réaliste avec des émotions. C'est ce que cherche à faire **Emoshape** (2014, USA, $370K), créée par le Français Patrick Levy-Rosenthal avec son composant électronique Emotion Processing Unit (EPU, destiné à déterminer en temps réel les émotions des utilisateurs et à permettre aux robots et autres applications comme des jeux de répondre avec un état émotionnel en phase avec celui de l'utilisateur (explication).

Le chipset récupère les informations de bas niveau issues de diverses sources d'informations comme les analyses du visage réalisée par Affectiva, des analyses de la voix réalisées par d'autres outils et d'autres informations issues de capteurs divers (pouls, ...) et permet à une IA interagissant avec l'utilisateur d'adopter son propre état émotionnel, sur une palette riche de 64 trillions d'émotions différentes (vidéo), que ce soit par de la parole de synthèse comme avec WaveNet de DeepMind, de la génération d'avatars ou même la gestuelle dans le cas d'un robot humanoïde.

Le système s'enrichit de plus par l'apprentissage pour développer des états émotionnels associés aux utilisateurs qui interagissent avec lui.

[635] Voir Trint's AI-powered plug-in automatically creates captions for Premiere Pro CC, mai 2018.

Il peut par exemple être associé à un générateur de langage naturel pour lui permettre d'accentuer son intonation en fonction des interactions émotionnelles avec l'utilisateur et des textes générés par l'IA. Le chipset peut être exploité dans divers contextes : robots, enceintes vocales, jeux vidéos, etc. C'est de l'*emotion in a box*. Et ce n'est qu'un début. Ce genre de composant ou les fonctions associées seront peut-être un jour directement intégrés dans nos smartphones et laptops et leurs logiciels tiendront compte de nos émotions pour interagir avec l'utilisateur. On peut par exemple imaginer comment un moteur de recherche tiendrait compte de nos états émotionnels pour ajuster ses résultats. Histoire par exemple de ne pas vous déprimer plus si vous l'êtes déjà !



Diffusion

Newswhip (2011, Irlande, $9,1M) propose un outil d'analyse de l'écho des médias et sujets dans les réseaux sociaux. Il permet d'affiner sa stratégie rédactionnelle pour que les sujets publiés collent bien aux attentes des lecteurs. Ils sont utilisés par des médias anglo-saxons comme le Huffington Post, Buzzfeed, la BBC et The Guardian.

NEWSWHIP

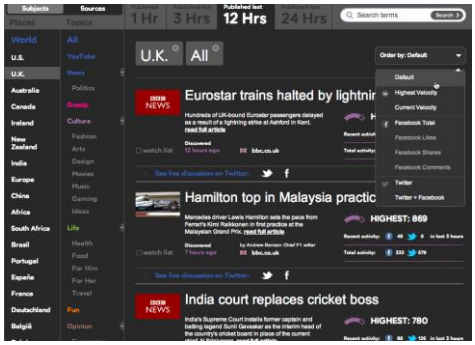
analytics de l'écho des médias et sujets dans les réseaux sociaux

réalise des prévisions de sujets à fort écho

à base de machine learning

références : Huffington Post, BuzzFeed, BBC, The Guardian

2011, USA, $9,1M



C'est aussi ce que fait **Banjo** (2010, USA, $121M) et qui ne cible pas que le marché des médias, sinon, il n'aurait pas levé ce montant énorme.

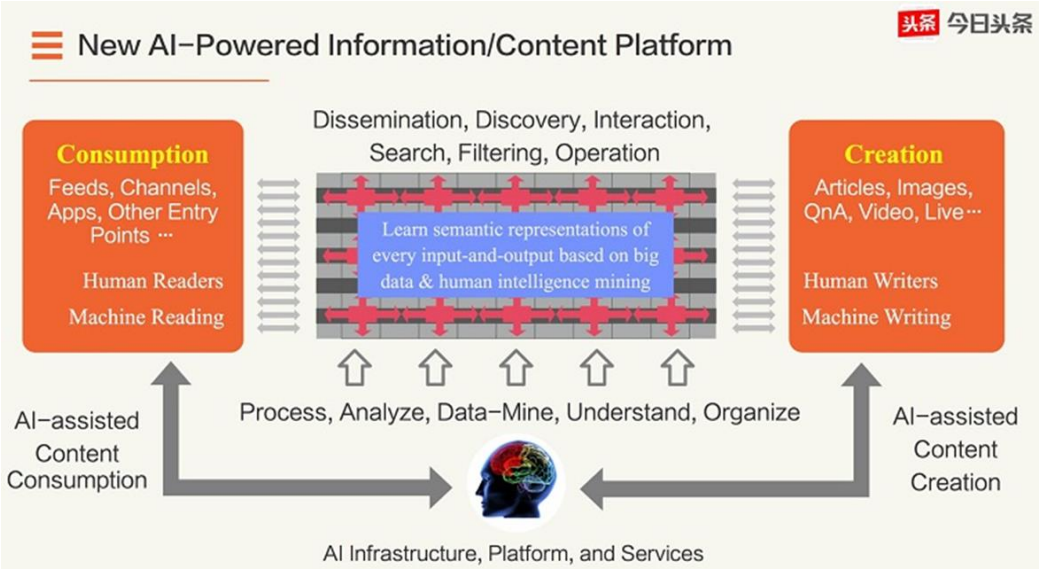
La startup **Echobox** (2013, UK $3,4M) propose de son côté Larry, un assistant dédié à la diffusion des contenus de médias dans les réseaux sociaux exploité par le Monde, Le Figaro, Libération, VICE et New Scientist. Comment ça marche ? Leur IA analyse les contenus du média et les tendances dans les réseaux sociaux, puis pousse ces contenus dans la page Facebook (ou autre) du média en générant automatiquement les titres, résumés et illustrations, histoire de maximiser leur diffusion. Ca ne va pas jusqu'à choisir les illustrations pour les publier sur Instagram en fonction des photos qui sont populaires dans ce service. C'est la prochaine étape !

La recommandation de contenus musicaux ou vidéo est un gros sujet pour le machine learning et l'IA. **Netflix** améliore de manière continue ses algorithmes pour faire consommer des contenus à ses abonnés (et aussi, cacher indirectement la pauvreté de son catalogue).

La R&D de la **BBC** expérimentait en 2018 une IA pour optimiser la programmation d'archives vidéo sur la chaine BBC 4[636]. **IBM Watson** est aussi appelé à la rescousse pour faire de la recommandation de vidéos sans qu'il soit évident d'évaluer les progrès en la matière[637]. Le système analyse les vidéos que l'utilisateur a consommées au niveau audio, vidéo et métadonnées pour faire des recommandations. Le problème avec ce genre de systèmes est qu'ils ne peuvent jamais récupérer ces informations sur tout ce que l'on a pu voir depuis sa naissance. Et le risque est que si vous aimez trop les films d'actions, il ne vous proposera jamais de voir Amélie Poulain qui aurait tout de même pu vous plaire. C'est le biais du rétroviseur, une fois encore !

Enfin, cela fonctionne aussi dans la musique, comme chez **Spotify** ou **Pandora**. Et chez **Decibel Music Systems** (2010, UK) qui fournit cela son application MusicGeek pour faire de la recommandation. Et qui utilise IBM Watson.

Toutiao (2012, Chine, $3,1B) est un OVNI dans la sphère des startups de l'IA. Cette startup a en effet battu des records avec son financement de $3,1B et sa valorisation de $20B. Tout cela pour un aggrégateur de news !



La startup est en fait une filiale du groupe ByteDance qui ne couvre que le marché chinois. Toutiao s'appuie sur de l'IA pour sélectionner les articles et vidéos à mettre en avant pour chaque utilisateur, pour réécrire les titres des articles afin d'améliorer leur référencement et taux de click et même pour en écrire, sur le sport. Leur audience est de 120 millions de Chinois. Ils sont aussi tentés par l'international[638] et par la création d'un chipset d'IA sur mesure.

Monétisation

D'une manière générale, l'IA peut aider les médias à identifier les sujets porteurs en analysant les tendances dans les médias sociaux et à agencer le sommaire des médias dans leur version web et mobile.

Le californien **True Anthem** (2008, USA, $11,2M) est une plateforme intégrée de distribution de contenu destinée aux médias. Elle permet notamment l'optimisation de la distribution des contenus au travers des médias sociaux, via un ciblage de contenus assisté par IA, qui décide notamment du moment optimum pour publier les contenus.

---

[636] Voir BBC Four announces experimental AI and archive programming, août 2018.

[637] Voir Watson Media Video Recommendations.

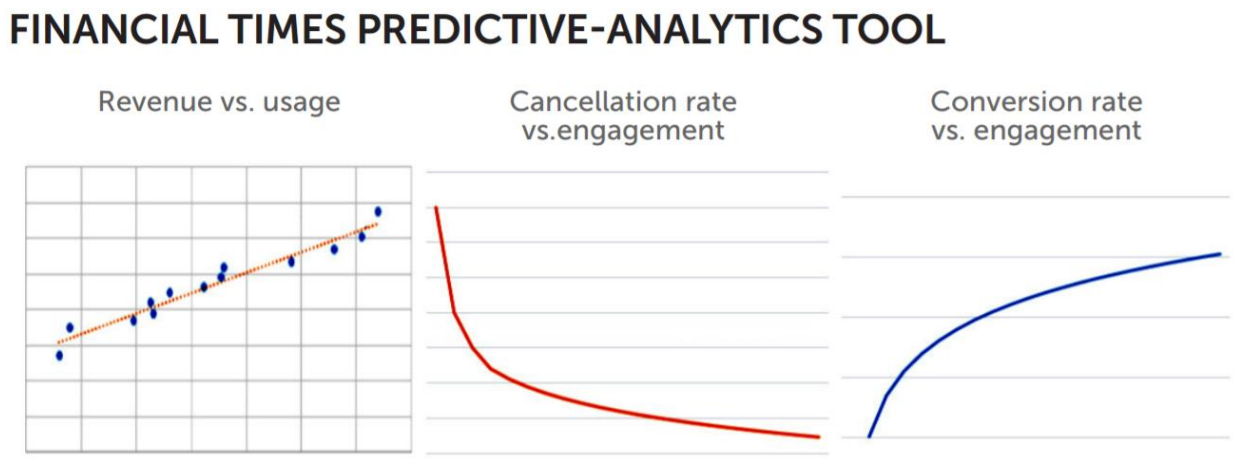
[638] Toutiao a fait notamment l'acquisition de l'éditeur d'applications mobiles **News Republic** (2013, France, $10,3M) en novembre 2017.

Leur service est exploité par Reuters et CBS Interactive. D'un point de vue technique, True Anthem a l'air d'exploiter des systèmes d'analyse du langage (NLP) et des moteurs de règles.

Adomik (2012, USA, $1,3M) est une startup française qui propose un outil de prévision à base de machine learning pour optimiser la publicité programmatique. L'outil est surtout destiné aux publishers.

De son côté, le **Financial Times** utilise une solution à base de machine learning pour déterminer la corrélation comportementale entre l'engagement dans le média et le churn[639]. Le **Wall Street Journal** analyse son lectorat en ligne pour proposer des offres d'abonnement payant sur mesure. Une sorte de freemium automatisé par le machine learning qui exploite une soixantaine de paramètres de comportement des lecteurs[640].

Enfin, **Red Points** (2012, Espagne, $26,2M) détecte les copies illicites de contenus en ligne et c'est loin d'être le seul à faire cela.



The Financial Times predictive analytics tools correlate data about revenue, content usage, cancellation rates, engagement, and conversion.

Internet

Les solutions sur Internet font évidemment abondamment appel aux techniques de l'IA et nous en avons déjà cité un bon nombre. L'une des plus connues est la mécanique de recommandation de contenus vidéo de Netflix qui s'est améliorée sur une période de plus de 10 ans, et exploite différentes techniques de machine learning.

En voici quelques-une qui sont un peu hors catégorie. Cette catégorie est évidemment bien plus large que cela.

Nudge.ai (2014, Canada, $4M) propose un plugin pour Chrome qui analyse vos emails et autres communication (limitées a priori au navigateur) pour vous proposer avec qui rentrer en relation. Cela serait adapté aux commerciaux et concurrencerait les solutions Einstein de Salesforce qui ont une finalité voisine. En fait, la solution s'intègre avec Salesforce. L'histoire ne dit pas quelle technique d'IA est exploitée dans l'outil. Probablement des briques de traitement du langage. C'est commercialisé de 0 à $60 par mois selon le niveau de fonctionnalités.

[639] Source : https://digitalcontentnext.org/blog/2017/06/13/artificial-intelligence-gains-momentum-news-media/.

[640] Voir How AI powers the Wall Street Journal's dynamic paywall de Laurie Clarke, octobre 2018.

Le site de rencontres **eHarmony** (1999, USA, $121M) utilise le machine learning pour améliorer le matching proposé, ce qui est très romantique mais fonctionne peut-être.Ont-ils accès à un historique qui va juste qu'aux prononciations de divorces ? Il faudrait en disposer pour faire du prédictif bien documenté[641] !

Tourisme

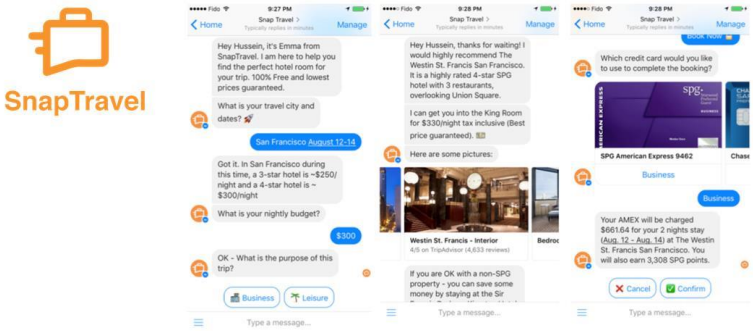
Le tourisme est un autre terrain de jeu propice aux innovations à base d'IA : les données sont abondantes, notamment via les intermédiaires de la réservation en ligne et chez les compagnies aériennes et leurs services de réservation mutualisés tels que **Sabre** et **Amadeus**[642]. C'est aussi un marché très grand public qui peut exploiter les outils de la mobilité et ceux des objets connectés.

Les systèmes de réservations de voyages exploitent toutes les techniques imaginables de « yield management » pour les remplir aux prix les plus élevés. Certains systèmes exploitent de la logique floue, d'autres du machine learning. **PriceMoov** (2016, France, 3M€) optimise ainsi les tarifs de différents services comme la location de véhicules (Rentacar), de chambres d'hôtel, de location d'équipements ou de billetteries en ligne. Leur "Dynamic Pricing" propose des prix dynamiques permettant d'optimiser le CA en s'appuyant sur l'état de la demande, les prix pratiqués par la concurrence et aussi la météo. Le tout exploite la solution logicielle de **Dataiku**. Il n'est pas sûr que cela améliore la satisfaction des clients finaux pout autant !

Chatbots

Les chatbots de préparation de voyages sont très nombreux, surtout aux USA. On en trouve qui sont attachés à des niches variés ou à des offres spécifiques comme pour les trains **Amtrak**, pour le processus de checking en ligne de **KLM**[643], **Air France KLM** qui a un chatbot pour Facebook Messenger, avec le chatbot de **Voyages SNCF** sur Facebook Messenger (vidéo), dans le groupe **Accor Hotels,** toujours sur Facebook Messenger[644] pour gérer les réservations dans un millier d'hôtel, soit le quart des hôtels du groupe, le tout étant couplé à un système de « Smart Pricing » (yield management), ou **Ask Mona** (2017, France), un chatbot de sélection de visites culturelles en France.

SnapTravel (2016, USA, $9,2M) permet de choisir son hôtel en fonction de ses contraintes budgétaires et via divers supports de communication (SMS, Facebook Messenger et même Slack). Il associe comme certains chatbots de l'IA et de l'intervention humaine et scanne les offres d'Expedia, de Priceline et de dizaines de sites.



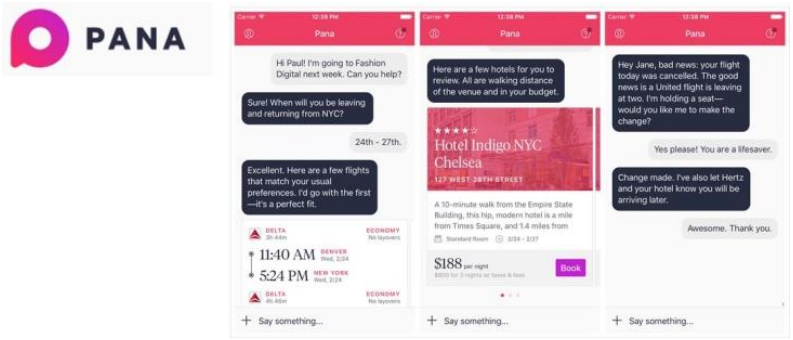
[641] Voir eHarmony: How machine learning is leading to better and longer-lasting love matches, mars 2018.

[642] Voir ce qu'ils en pensent dans Defining the future of travel through intelligence d'Amadeus, 2017 (22 pages).

[643] Voir cette liste de quelques dizaines de chatbots : https://www.30secondstofly.com/ai-software/ultimate-travel-bot-list.

[644] Que j'ai testé et qui n'apporte pas grand-chose, et en plus est très lent.

Bref, c'est un moteur de recherche à commande textuelle.

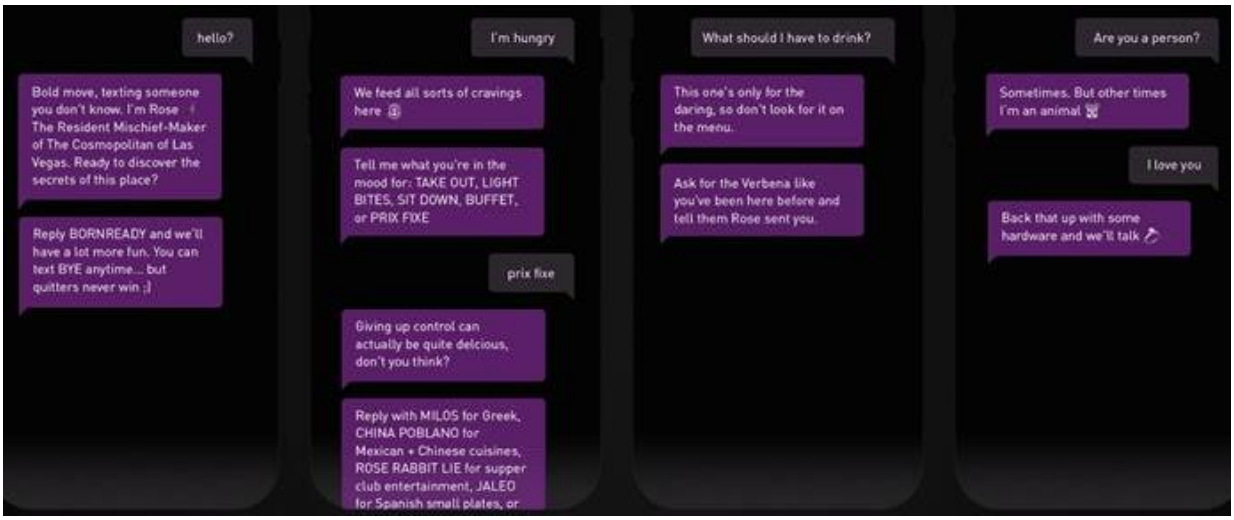


Pana (2015, USA, $1,5M) est un équivalent destiné aux voyages professionnels. **TravelAppeal** (2014, Italie, $7,1M) est un chatbot pour Facebook Messenger pour gérer son séjour en temps réel et obtenir des réponses aux questions courantes pendant le voyage.

Le chatbot s'alimente en aspirant les contenus de 500 sources Internet dont les réseaux sociaux et sites liés au voyage et au tourisme.

Le même principe est en place depuis début 2017 dans l'hôtel **Cosmopolitan** de Las Vegas qui utilise son chatbot Rose (*ci-dessous*) fonctionnant notamment via SMS qui permet de passer toutes sortes de commandes. Ce concierge autonome aurait permis d'augmenter de 39% les commandes d'extras par les clients et 90% d'entre eux y feraient appel.

L'hôtel qui est relativement récent comprend 3000 chambres. Le chatbot a été développé par l'agence américaine **R/GA**. Vous pouvez tester le chatbot en appelant le + 1 702 930 8188 !



Les applications mobiles généralistes chatbot de préparation de voyage ne manquent pas aux USA. Nous avons par exemple **Lola Travel Company** (2015, USA, $44,6M), créé par Paul English, le fondateur du moteur de recherche de voyages **Kayak** revendu à Priceline pour $2B**, Mezi** (2015, USA, $11,8M) et **Skyscanner** (2003, USA, $197M, acquis par le chinois **Ctrip** en novembre 2016). Toutes ces applications se ressemblent et accèdent généralement aux mêmes sources de données[645].

On commence aussi à voir des hôtels installer des systèmes à commande vocale dans les chambres. C'est le cas au **Encore** et au **Wynn** à Las Vegas dont les chambres sont toutes équipées de systèmes **Amazon** Echo pour piloter, essentiellement… la lumière. La commande vocale peut aussi servir à effectuer sa réservation, cette fois-ci à partir de chez soi, avec ces enceintes connectées à commande vocale dont le taux de pénétration dans les foyers dépasse maintenant celui des weareables, aux USA.

[645] Voir Artificial Intelligence (AI) in the Travel Industry, avril 2017.

Mais les statistiques d'usage semblent montrer que, pour l'instant, les consommateurs sont encore hésitants à passer des commandes via la voix et Alexa/Echo, préférant le faire dans des sites web ou applications mobiles, qui permettent de plus facilement vérifier que l'on ne s'est pas trompé. C'est peut-être aussi lié à un phénomène d'acoutumance lent à se produire comme cela a été le cas sur le commerce en ligne aux débuts de l'Internet.

Parcours touristiques

La création de parcours touristiques personnalisés devrait être un bon champ d'application de l'IA. On indiquerait sa ville, ses préférencs en termes de types de visite, le nombre de jours, les moyens de transports préférés et le budget et le système produirait automatiquement des propositions d'agendas avec horaires, transports et un forfait pour tout payer.

Mais aujourd'hui, c'est encore plutôt du domaine de la science fiction et de la singularité. Pourquoi donc ? Parce qu'il est très difficile d'obtenir toutes les données structurées nécessaires, que les marchés sont fragmentés, que les billetteries en ligne ne sont pas standardisées pour les visites et qu'il faut tout refaire dans chaque ville. Mais cela arrivera bien un jour. Vous objecterez avec tel ou tel service qui existe déjà, mais vous rendrez compte, de près, qu'il manque toujours quelque chose comme la création d'un parcours qui intègre les temps de transport.

Mais l'application mobile **Google Trip** qui fonctionne même en mode déconnecté commence à s'en approche tout de même (vidéo). Mais la vidéo de présentation est trompeuse ! Les parcours de visite sont préformatés et ne sont pas personnalisables.

Wayblazer (2014, USA, $8,7M) propose des recommandations contextualisées et personnalisées et tourne au-dessus d'IBM Watson (vidéo). Il personnalise l'accompagnement photographique des propositions en fonction des recherches textuelles multicritères de l'utilisateur.

C'est une sorte de concierge numérique commercialisé aux professionnels du tourisme. Une solution équivalente est proposée par **GoMoment** (2010, USA).

Gogobot (2010, USA, $39M) et son application Trip.com utilise un modèle prédictif qui exploite la segmentation socio-démographique du voyageur, le moment et la météo pour proposer des visites. Mais l'intégration n'est pas extraordinaire au premier abord lorsque l'on teste le site qui sépare hôtel et avion alors que l'offre devrait être intégrée, comme dans **Opodo**. La startup a été également acquise par le Chinois **CTrip**.

Et des guides de visites en réalité augmentée, qui seraient des équivalents de **Pokemon Go** servant à quelque chose ? Cela arrive au compte goutte, mais avec des couts de production par attraction qui sont encore trop élevés pour être généralisés. Reste aussi à inventer une IA qui rendrait les serveurs des cafés parisiens plus sympas et orientés clients !

Expérience touristique

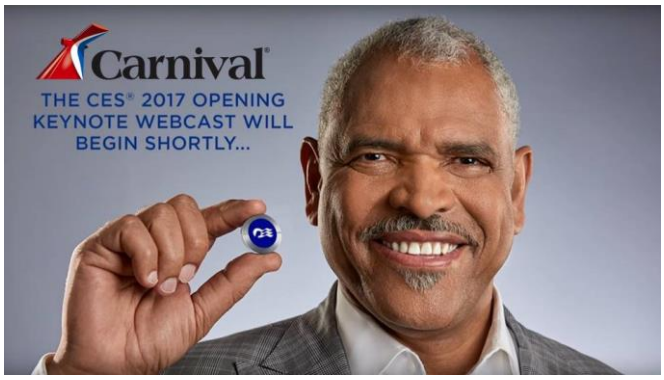
L'expérience touristique peut s'améliorer en tirant parti de l'IA à différents étages.

J'ai pour l'instant découvert cet outil de prévision proposé aux hôteliers par la startup française **Victor&Charles.**

Elle s'appuie sur IBM Watson et exploite toutes vos données publiques des clients disponibles dans les réseaux sociaux pour en analyser l'influence, les affinités et l'humeur. Il propose alors des recommandations à l'hôtel qui vous accueille pour lui permettre de personnaliser votre arrivée, et notammentt de trouver la personne la plus appropriée pour s'en charger.

Cela s'appliquera plutôt à des hôtels quadri-étoilés ou plus ! La solution utilise IBM Watson Conversation, Natural Language Understanding, Personality Insights et Tone Analyzer. Le matching de personnalité est une fonctionnalité que la startup souhaite commercialiser au-delà du marché de l'hôtellerie. Son développement n'a pas duré plus de deux mois.

Et puis nous avons cette expérience client en environnement fermé présentée par **Carnival** lors d'un keynote au CES 2017 (vidéo). Elle consiste à proposer un badge RFID aux passagers des paquebots qui permet d'accéder à tous les services du navire, ces services étant personnalisés en fonction des préférences et de l'activité des passagers. Le tout avec force machine learning exploitant l'historique de consommation des passagers.



Depuis quelques années, vous pouvez installer sur votre smartphones diverses applications, notamment de **Google**, qui traduisent automatiquement la signalétique tout comme les menus de restautants.

Robots

Si vous aimez les robots, le tourisme pourra vous donner l'occasion d'en croiser, mais plutôt rarement, et en Asie, le continent qui n'a pas peur des robots.

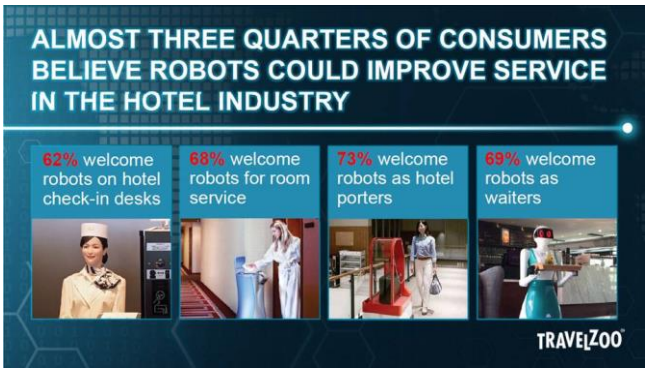
Tout d'abord en vous équipant vous-mêmes d'une valise robot comme l'étonnante **Cowarobot** (vidéo). D'origine chinoise, la startup avait réussi sa levée de fonds sur IndieGogo avec $581K de récoltés en septembre 2016. A ce jour, les valises ne sont toujours pas livrées aux early-adopters !

Vous pouvez aussi croiser des robots mobiles d'information de **Qihan** dans les aéroports comme à Shanghai. **Starwood** a mis en service un robot majordome à roulettes et tablette dans un hôtel à Palo Alto en 2014. Il permet de livrer dans les chambres les petits ustensiles demandés par les clients, comme des serviettes, des brosses à dent ou des accessoires électriques. Il provenait de **Savioke** (2013, USA, $34M) qui est de la même région. Ce robot est aussi déployé dans des hôpitaux. Le cours de l'Histoire aurait été changé si un tel robot avait été installé dans le Sofitel de New York en 2011 !

Enfin, si vous allez au Japon, vous pourrez faire un séjour dans l'hôtel pilote **Henn-na** de Tokyo avec ses 75 chambres et dont l'accueil est réalisé par des robots depuis 2015 (vidéo).

Vous le choix entre une hôtesse robot ou un vélociraptor robot qui ne font que servir d'interface visuelle pour l'automate qui vous permet de faire votre checking et qui existe déjà dans diverses chaines d'hôtel en France. Ce même hôtel robotise le transport de vos bagages dans votre chambre.

Dans des cas relevant moins de la science fiction, certains hôtels commencent à installer des bornes de check-in avec reconnaissance faciale.



Ce n'est qu'un début ! Après, si vous avez besoin de rencontrer de vraies personnes pour alimenter votre organisme en sérotonine et en dopamine, c'est une autre affaire !

Mode et luxe

Les marchés de la mode et du luxe sont hétérogènes mais partagent un certain nombre d'aspects pour ce qui est des cas d'usages de l'IA. Ils sont, pour une part, voisins de ceux du retail, du fait des réseaux de point de vente des marques concernées. Les marques grand public peuvent aussi adopter des chatbots pour fluidifier leur relation clients.

Les sites web peuvent être enrichis de fonctions de recherche de produit par similarités visuelles. L'IA peut aussi intervenir dans le processus d'étude de tendances, dans la création ainsi, en aval, que dans la détection de contrefaçons.

La présence d'un chercheur en IA dans une publicité de parfum d'Yves Saint Laurent est anecdotique mais symbolique de la valorisation de ce genre de profil professionnel. Ces secteurs d'activité sont très traditionnels. Ils s'appuient sur des jeux de données de volume variable selon le niveau d'élitisme des produits commercialisés. Le fashion grand public est un marché de volume fragmenté. Le luxe est tout aussi fragmenté mais génère moins de volume.

Création

L'IA peut être impliquée dans la conception de produits ainsi que dans les unités de production, surtout celles qui sont industrielles.

Cela peut aller jusqu'à la création de produits avec l'aide d'IA. En complément des outils de conception 3D (CAO), le principe consiste à utiliser des réseaux de neurones génératifs (GANs) pour créer des modèles à partir de modèles existants. Le débat sur la créativité de l'IA va bon train dans un pareil cas. En pratique, les créatifs humains jouent avec la palette qui est proposée par les logiciels et retiennent ce qui leur semble le plus intéressant. Ces IA restent des outils puissants à disposition des créatifs plus que des outils remplaçant les créatifs. Comme un traitement texte pour un auteur ou un logiciel de CAO pour un ingénieur.

L'un des exemples les plus connus est l'artiste et chercheur en IA Robbie Barrat de **Balenciaga** qui s'est ainsi appuyé sur une IA générative pour créer une collection complète s'appuyant sur des modèles existants de la marque[646].

L'IA créait des variantes de styles et couleurs à partir de l'existant. Le résultat est assez brouillon[647].

Est-ce qu'une IA peut détecter les tendances en amont de leur apparition via une détection de signaux faibles[648] ? C'est une promesse facile à faire mais plutôt difficile à réaliser. Les bases d'entraînement ne sont pas évidentes à constituer. Les signaux faibles sont de nature diverse et pas évidents à intégrer. Qui plus est, le principe même de la création et du lancement de tendances consiste à aller dans des zones non explorées. S'appuyer sur trop d'IA pour détecter des tendances pourrait amplifier « l'effet du rétroviseur ». Des outils d'IA ou de représentation des tendances peuvent servir à comprendre l'environnement dans lequel évolue une marque. C'est ce que propose **Heuritech** qui analyse les tendances de la mode dans des bases d'images hétérogènes, avec les couleurs et motifs qui plaisent.

L'IA peut aussi servir à créer de nouveaux parfums. Plusieurs projets vont dans cette direction comme celui d'**IBM Research** et **Symrise**, un producteur de parfums qui fournit Estee Lauder et L Brands.

Leur AI dénommée Philyra s'appuie sur du machine-learning pour exploiter la base de données de Symrise qui comprend 1,7 millions de formules et identifie les "oceans bleus" de formules non testées ou non utilisées (vidéo). Le système suggère alors des nouvelles combinaisons d'essences et de parfum qui pourraient plaire à certains segments de clientèle[649].

Edited (2009, UK, $6M) est un outil d'aide à la décision de création de gammes. Il permet aussi de réaliser une analyse concurrentielle, via l'analyse de défilés de modes et de sites de e-commerce.

Production

Le marché du textile peut aussi faire appel à des robots, comme ceux de **Softwear Automation** (2007, USA, $7,5M) qui sont capables de fabriquer des chaussures, des t-shirts et des matelas (vidéo). Ce sont surtout des automates qui répètent les mêmes gestes à longueur de journée mais ils disposent tout de même de systèmes de vision artificielle pour gérer leur production.

Personnalisation

Dans l'habillement, la personnalisation des produits via des scanners 3D, de la reconnaissance d'images prises avec des systèmes de captation 3D en boutique ou des smartphones permettent de déterminer les mesures exactes et/ou la corpulence du client ou de la cliente. Les clients se voient alors proposer les bonnes tailles de vêtement.

---

[646] Voir Can an AI do Balenciaga better than Balenciaga?, septembre 2018.

[647] Voir https://threadreaderapp.com/thread/1021476460345708544.html. Notons à contrario que la génération de peintures artificielles fonctionne assez mal si l'on cherche à obtenir un rendu réaliste. C'est le cas avec AI is used to automatically create nude portraits.. and the results are horrifying, mars 2018, qui évoque le cas de GANs générant des peintures de nus.

[648] C'est le propos de l'article Intelligence artificielle et luxe : l'alliance des possibles de Laura Perrard, mai 2018.

[649] Voir Breaking new fragrance ground with artificial intelligence (AI): IBM Research and Symrise are working together, octobre 2018. La solution rappelled aux industries traditionnelles qu'il est bon d'avoir un système d'information qui capte bien l'historique métier de la société.

Ce genre de méthode intervient aussi de plus en plus dans le calibrage des montures de lunettes aussi bien pour la vente en opticien qu'en ligne. Là encore, avec de la reconnaissance visage et des éléments clés dans les visages.

La marque **Stitch Fix** utilise un « styliste digital » paramétré par les clients avec leurs goûts et préférences vestimentaires, leur mode de vie, loisirs, choix de looks. Le tout récupéré notamment via leur compte Pinterest. L'IA propose ensuite des choix vestimentaires correspondant à ce profil.

La marque **Eison Triple Thread** produit des costumes sur mesure avec sa solution FITS en s'appuyant pour sa part sur les goûts musicaux du client qui sont détectés sur sur compte Spotify et associés à un questionnaire de profiling voisin de celui de Stitch Fix[650]. Le système génère alors des styles de costumes alignés sur ces goûts. In fine, c'est toujours le client qui choisit.

Expérience client

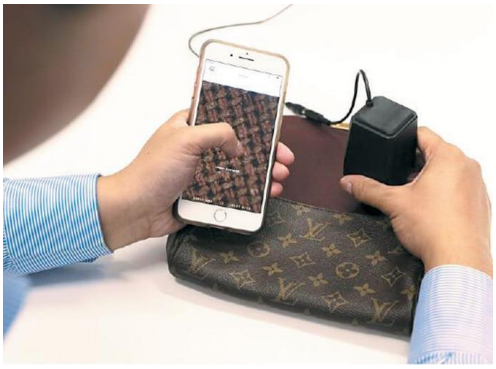
L'IA peut servir à trouver des produits similaires dans les catalogues en ligne et, dans la lignée, identifier des contrefaçons ou produits gris, au niveau des produits eux-mêmes comme à celui du packaging.

Farfetch (2008, UK, $705M) est un site de e-commerce et une place de marché dans la mode et le fashion. Il s'appuie sur de l'IA pour améliorer les parcours clients avec la plateforme de **Certona** (2004, USA, $37M).

On peut peut évaluer l'émotion « extérieure » des clients en les détectant sur les visages captés par des caméras dans les boutiques ou sur mobiles et laptops. C'est ce que proposent des sociétés comme **Affectiva** ou **Datakalab** qui analysent le visage d'une ou plusieurs personnes pour capter différents types d'émotions affichées. Cela peut être exploité avec de la vidéo surveillance de boutique pour évaluer la performance d'une devanture ou de contenus d'affichage dynamique.

Contrefaçons

La détection de contrefaçons est un objectif clé de nombreuses marques de luxe, qu'il s'agisse des sacs de Louis Vuitton, des parfums ou dans la joaillerie. Celle-ci concerne aussi bien les produits que les packagings.



En plus de **Data & Data**, déjà évoqué dans le reste du document, nous avons aussi **Entrupy** (2012, USA, $2,6M) approche la recherche de contrefaçons de manière originale avec une petite caméra portative photo permettant de prendre des photos en mode macro des objets et d'analyser leur matière avec une IA. La solution serait fiable à 99,1%, ce qui est plausible si la base d'entraînement est bien alimentée. Et pour cause, puisqu'elle contient plus de 30 millions d'images (vidéo). Une technique équivalente serait capable de détecter le coup de peinceau de peintres dans 80% des cas et d'identifier des faux[651].

Cypheme (2015, France/USA, $1,3M) est autre une startup dans ce secteur, qui propose une application mobile de détection de produits contrefaits, s'appuyant sur un algorithme de machine learning appliqué à la qualification d'images, associé à une étiquette spéciale. C'est une sorte de Shazam de la contrefaçon.

---

[650] Voir Fashionably AI - Eison Triple Thread reimagines suit-making with music and artificial intelligence, de Henry Pickavet, août 2018.

[651] Voir AI Used To Spot Art Forgeries By Looking At Brushstrokes de Tyler Lee, novembre 2017.

Services et conseil

Les cabinets de conseil, d'études de marché, analystes, services et autres consultants indépendants sont les premiers à secouer le marché dès qu'une nouvelle technologique apparaît. Ils pratiquent à outrance le marketing de la peur, expliquant aux entreprises qu'elles doivent à tout prix adopter ces nouvelles technologies sous peine de se faire disrupter ou uberiser en cas d'immobilisme technologique[652].

Ces métiers représentent plus de 400 000 emplois en France rien que dans le conseil et la réalisation de projets informatiques !

Mais ces cordonniers sont-ils bien chaussés ? Est-ce que ces « travailleurs du savoir » sont bien outillés avec l'IA pour être performants et pertinents ? Au nez, pas vraiment. Les méthodes de travail des consultants et des analystes n'ont pas beaucoup évolué, avec ou sans IA. Mais il y a du potentiel et les idées ne manquent pourtant pas. C'est un monde d'abondance de *couldashouldawoulda*[653].

Pourquoi en est-on à ce point ? La faible automatisation d'un métier a plusieurs origines : sa grande fragmentation organisationnelle, la diversité des tâches et des situations, la déstructuration des données utilisées, le fait que l'Internet soit un gros foutoir mal organisé et moins ouvert qu'il n'y parait, l'absence de standards et les limites actuelles des techniques logicielles et de l'IA.

Qui plus est, le marché de l'outillage du conseil est aussi un marché relativement limité. Lorsque l'on additionne la fragmentation des besoins et celle du marché, cela donne un marché faiblement innovant côté outillage. Sauf lorsque ce marché peut réutiliser des outils conçus pour des marchés plus grands.

Passons maintenant en revue quelques-unes de ces applications potentielles de l'IA dans différents métiers du conseil en passant par le conseil en stratégie et en management, la gestion de projets, les études de marché et les projets informatiques.

Conseil en stratégie, innovation et management

Le conseil en stratégie, innovation et management démarre souvent par une phase de découverte de l'entreprise cliente avec un grand nombre d'entretiens avec ses collaborateurs et même avec d'autres entreprises. Cette tâche peut être facilement fastidieuse car il faut généralement conserver une trace écrite de ces entretiens, puis en faire une synthèse qualitative et quantitative[654].

C'est là que peuvent être exploités des outils de transcription automatique des entretiens s'ils sont enregistrés. Le speech-to-text est une technique encore imparfaite mais elle peut permettre de gagner du temps. Elle est même intégrée dans certains outils bureautiques comme **OneNote** de Microsoft.

La synthèse automatique peut être réalisée par l'extraction automatique d'éléments clés dans le texte. Cela peut concerner des expressions ou des mots qui sortent de l'habitude, ou par la détection de sentiments dans les textes. Des statistiques des mots clés employés peuvent aussi être réalisées.

---

[652] Le message le plus classique consiste à mettre en regard la valorisation boursière de sociétés du numérique (Uber, AirBnb, Facebook, Google) et celle d'entreprises traditionnelles (automobile, BTP, hôtellerie), en négligeant le chiffre d'affaires et la profitabilité. Bref, cela consiste à comparer des choux et des carottes. Si toutes les entreprises étaient des « pure players » numériques, on serait bien avancé et incapables de se loger, de s'alimenter, de se chauffer, d'accéder à de l'eau potable et de se déplacer. La comparaison n'est pas la bonne. Toutes les entreprises qui sont dans les biens de consommation ou infrastructures matérielles n'ont pas forcément vocation à quitter ces métiers, même si cela peut parfois arriver, comme la manière dont le groupe Accor Hôtels a séparé l'activité de l'hôtellerie de celle de l'immobilier associé. Bref, la « transformation digitale » ne doit pas prendre comme repère les pure players de l'Internet.

[653] Comme évoqué dans L'intelligence artificielle au service des cabinets de conseil, par Matthieu Gufflet, CEO d'Epsa Groupe et Thomas Bourgeois, CEO de Dhatim, juillet 2018, qui fait la promotion de la notion de « consultant augmenté ».

[654] Voir Le conseil de l'innovation mise sur l'intelligence artificielle, 2018.

C'est ce qu'avait fait la société **Proxem** pour l'équipe En Marche en 2016 en amont du lancement de la campagne présidentielle d'Emmanuel Macron.

Ces analyses recouvrent parfois une analyse de la concurrence qui passe par des entretiens et l'exploitation de sources ouvertes ou fermées qu'il faut consolider. Il est notamment critique de récupérer des indicateurs économiques quantitatifs sur ces concurrents. L'IA intervient peut dans ce processus à ce jour.

Dans les phases d'idéation, des de représentation des connaissances peuvent être exploités avec des arbres de décision, des métaplans virtuels créés sur tableaux de bord interactif. Mais là encore, sans forcément faire appel à de l'IA. Elle pourrait cependant intervenir pour réaliser automatiquement des regroupements syntactiques ou thématiques de diagnostics ou de propositions.

Suit en général la production de rapports et de présentation. Leur efficacité peut être améliorée avec des outils d'analyse du langage : corrections, calculs d'indices de lisibilité et analyse d'efficacité d'argumentaires. Il est bien plus facile d'automatiser les prestations de services sur des métiers opérant sur des données quantitatives comme la finance, la comptabilité, les achats, les facturations, les audits et le marketing.

Gestion de projets

C'est une activité courante dans les cabinets de conseil et les entreprises de services numérique (ESN). Les outils traditionnels utilisés servent à la gestion des plannings et des ressources. Ce sont des gestionnaires de projets du marché qui peuvent intégrer des briques d'IA selon les cas de figure.

Ils peuvent intégrer des fonctions prédictives et identifier à l'avance les zones de congestion ou de retards potentiels des projets, histoire de les prévenir en allouant les ressources de manière appropriée. La question clé étant de disposer de données d'entraînement suffisantes pour que les IA se débrouillent.

Lili (2016, France) capte l'ensemble des données des projets dont les communications écrites entre intervenants pour prévenir ou gérer les litiges.

Des solutions à base d'IA peuvent aussi être mises en place pour le staffing des projets, pour trouver les bonnes compétences, vérifier la compatibilité humaine des équipes, composantes que nous avons partiellement évoquées dans la rubrique sur les ressources humaines.

Etudes de marché et sondages

Les études de marchés et sondages génèrent de gros volumes de données qui se prêtent à des analyses à base d'IA. C'est le métier de sociétés telles que Médiamétrie, BVA, Ipsos ou l'IFOP mais aussi le Gartner, Forrester ou IDC.

Les enquêtes terrains et sondages peuvent éventuellement faire appel à des chatbots textuels ou vocaux pour le questionnement des sondés. Cela a du sens économique mais peu biaiser les réponses générées.

Dans les baromètres d'audience de médias chauds comme la TV, des systèmes d'analyse du signal audio qui sort de la TV permettent d'identifier les programmes visualisés. C'est de l'IA ! D'autres solutions détectent automatiquement des logos dans les vidéos, pour évaluer la présence de marques et annonceurs dans les contenus à la TV tout comme dans les médias en ligne. L'analyse de sons et d'images en tout genre avec du deep learning devient un moyen clé d'auditer l'exposition médiatique et conversationnelle des marques.

De leur côté, les études de marché qualitatives impliquent la recherche, la compilation et la mise en forme de données hétérogènes issues de nombreuses sources. Cela comprend de la recherches bibliographiques sur Internet, l'usage de moteurs de recherche et de bases ouvertes ou payantes spécialisées.

Il y a encore beaucoup à faire pour automatiser la curation d'information. Les méthodes utilisées restent en général assez traditionnelles avec diverses astuces d'utilisation des moteurs de recherche, la recherche par mots clés dans des flux RSS, des systèmes d'alertes.

Ces techniques pourraient exploiter de l'IA pour faire des analyses syntaxiques, créer automatiquement des ontologies de domaines d'activité.

Quid d'un **Rapport du CES** qui serait généré automatiquement ou cet ebook ? Une bonne part de mon travail de rédaction relève d'un processus itératif. Découverte de nouveaux produits, recherche de leur description la plus complète et factuelle, recherche d'illustrations photos et vidéo, puis synthèse. Théoriquement, tout cela pourrait être automatisé. Il en va de même des comparatifs qui pourraient aussi l'être, par exemple entre tous les smartphones du marché. Pourant, ce qui est théoriquement possible ne l'est pas pratiquement ou économiquement car peu de professionnels ont ce genre de besoins.

On peut aussi faire appel à diverses bases de données d'entreprises et de startups qui consolident les informations publiques à leur sujet (comme Crunchbase, Skopai ou U-Change). L'analyse de vidéo et leur transcription automatique bien mise en forme pourrait aussi aider à trouver les informations pertinentes. La recherche d'information pourrait aussi bénéficier de l'existence d'outils d'élagage de textes, supprimant la rhétorique, les superlatifs, termes vagues et autres discours stéréotypés.

Ce genre d'activité implique aussi la consolidation de données chiffrées pouvant faire appel à des outils d'extraction de données chiffrées dans les textes. Je passe généralement pas mal de temps pour générer des tableaux avec les indicateurs économiques de quelques sociétés du monde du nuémrique. Cette compilation est encore bien trop manuelle. Et ne parlons pas des sociétés non côtées pour lesquelles les informations recherchées ne sont pas dans le domaine public ! Enfin, il faut produire des visuels et notamment des charts et autres business analytics. Et puis surtout, trouver des moyens de vulgariser certains contenus techniques. L'utilisation d'analogies ou d'exemples n'est pas encore à la portée des IA de traitement du langage ou de génération de résumés.

Des modèles mathématiques plus sophistiqués à base de machine learning pourraient aussi être développés pour croiser plusieurs études de marché, identifier les variations et écarts types, normaliser les études consolidées en fonction de leurs descriptifs qui sont souvent hétérogènes. Comme par exemple la comparaison des prévisions sur le pourcentage des emplois qui seront détruits et créés par l'IA dans les années à venir. Des outils de normalisation des études en termes de méthodologie, de périmètre, de champs et de timing seraient les bienvenues.

L'une des raisons pour lesquelles cette consolidation est difficile est la faible structuration des données publiées sur Internet. L'adoption des formats XML et RDF promus au début des années 2000 a été faible. Les initiatives de publication d'open data se multiplient par leur exploitation reste un processus très manuel.

Les études de marché ont plein de sources de données à leur disposition qui sont de plus en plus souvent exploitées par des progiciels à base d'IA. Que ce soit par exemple de Google Trends ou LinkedIn. Cela permet d'analyser les mouvements entre entreprises, les évolutions des compétences affichées dans les CV, les postes ouverts, la circulation des talents. Les outils ne reposent pas encore assez sur de l'IA. Ils pourraient y faire appel au minimum dans l'interprétation de données de langages non structurées.

D'autres sources permettent aux entreprises de mieux connaître leurs concurents. Leur CA, leurs clients, les gains et pertes de projets, la visivilité de leurs produits, les évolutions des équipes, des méthodes marketing, les verbatim marketing, la satisfaction clients, l'analyse de sentiments, celle des signaux faibles, les analyses temporelles, l'identification de leurs influenceurs. Voilà de quoi probablement alimenter quelques briques d'IA de traitement du langage et des données.

On rêve de moteurs de recherche d'un nouveau genre qui seraient capables d'interpréter nos questions et d'aller chercher tout seul les éléments de réponse sur Internet, les intégrer et les mettre en forme. C'est technique possible. Dans nombre de cas, il ne s'agit pas de créer des AGI (intelligences artificielles générales). Reste à le faire ! Qui se lance ?

Services informatiques

Terminons avec les services informatiques. Chez eux, qu'est-ce qui est répétable et automatisable ? Nombre de processus qui ne relèvent pas forcément de l'IA peuvent être autoamtisés avec de simples macros.

Dans l'IT, l'informatique d'infrastructures, les outils de monitoring commencent à exploiter le machine learning pour détecter les bizarreries et les tendances.

Est-ce que les développeurs remplacé par des IA ? Certaines actualités récentes le laissent croire mais c'est un mythe. Il me rappelle le débat sur les langages de quatrième génération dans les années 1980. Leur promesse n'a jamais été véritablement tenue même si l'on a pu bénéficier de progrès via la prolifération de « frameworks » associés aux grands langages de programmation du marché (Java, JavaScript, PHP, Python, …). Mais on a toujours autant besoin de développeurs. Tout simplement parce que la quantité de logiciels développés à augmenté, Internet et mobile obligent.

Education

Le secteur de l'éducation est depuis longtemps un grand champ de promesses pour les usages de l'intelligence artificielle. Les applications pressenties touchent souvent à la personnalisation de l'enseignement à distance via des agents conversationnels intelligents capables de suivre et accompagner pas à pas les élèves dans leur progression[655]. On en est encore au stade des expérimentations. La littérature sur le sujet est pour l'instant assez vague[656].

Fonctions	Plan cognitif	Plan Socio-affectif	Plan motivationnel	Plan métacognitif
Accueil et orientation	Informer sur le dispositif de formation	Initier la construction d'un sentiment d'appartenance	Faire émerger les objectifs personnels de l'apprenant	Inciter l'apprenant à faire le point sur ses stratégies cognitives
Organisation	Présenter les méthodologies appropriées	Réguler la dynamique de groupe	Accompagner le processus d'autonomie	Faciliter la planification de l'apprentissage
Pédagogie	Apporter des réponses ou les susciter. Remédier	Faciliter la collaboration des apprenants	Proposer des activités signifiantes	Susciter l'expression critique sur le dispositif
Socio-affectif Motivation	Personnaliser le soutien à l'apprentissage	Rompre l'isolement de l'apprenant	Lutter contre l'abandon	Faire prendre conscience de ses habiletés à collaborer
Technique	Aider à maitriser l'environnement d'apprentissage	Susciter l'entraide technique entre apprenants	Encourager l'utilisation des outils	Susciter la prise de distance réflexive sur les usages des outils
Métacognition	Faire conscientiser ses préférences cognitives	Faciliter la prise de conscience des états affectifs / tâches	Faire identifier les motivations intrinsèques	Inciter l'apprenant à apprendre à apprendre
Evaluation	Annoncer clairement les critères d'évaluation	Produire des rétroactions à portée formative	Encourager et féliciter	Aider à s'autoévaluer

En gris les interventions pouvant être confiées à un robot ayant été préalablement formé
En bleu, les interventions partagées par un robot et les tuteurs humains
En vert, les interventions ne pouvant être confiées qu'à des tuteurs humains

[655] La version la plus extrême de ces méthodes est le « neurohacking » des élèves qui consiste à décortiquer les processus de mémorisation des connaissances et à hacker le cerveau pour le faire fonctionner de manière optimale. Cela va jusqu'à faire des IRM fonctionnelles pendant des phases d'apprentissage pour identifier les parties du cerveau qui sont activées.

[656] En voici un exemple: The Future of AI and Education, mai 2018, qui n'est pas bien dense en études de cas !

Ces solutions sont difficiles à mettre au point et à déployer à grande échelle. L'éducation est un de ces secteurs d'activité où la fragmentation ralentit l'automatisation. Ici, il s'agit de la fragmentation des matières, contenus, méthodes et typologies d'élèves, multipliés par les variantes linguistiques. Comme les logiciels en général, l'IA est pertinente lorsqu'elle génère des économies d'échelle.

Dans Quelle place pour les robots dans le tutorat à distance ?, octobre 2016, Jacques Rodet pose de bonnes questions sur les usages du numérique et de l'IA dans l'enseignement. Le tableau *ci-dessous* décrit les tâches dans l'enseignement qui peuvent être automatisées et celles qui demanderaient encore des enseignants. Plus de la moitié des tâches requièrent ces derniers dans le tableau.

Applications

Voici quelques-unes des applications potentielles de l'IA dans l'éducation qui sont couramment citées dans la littérature sur le sujet voire expérimentées ou intégrées dans des offres commerciales[657] :

Notation automatique des élèves, ce qui n'aurait de sens que pour noter la partie qualitative de QCMs. Pour ce qui est de la notation de copies manuscrites d'élèves, les IA de reconnaissance d'écriture ne sont pas encore assez au point.

Support des élèves via un chatbot, ce qui suppose un environnement pédagogique bien formalisé dans le chatbot en question[658]. Le chatbot devrait aussi avoir accès à l'historique personnel de l'élève. Bref, savoir tout sur lui de sa scolarité et sur ses forces et faiblesses. Dans un monde parallèle ! Mais sur des tâches élémentaires, ce genre d'outil allégerait la tâche des enseignants, leur permettant de mieux jouer le rôle de mentor. Ces outils permettraient aussi d'accompagner les élèves hors des classes et dans les MOOC pour personnaliser l'enseignement, l'assimilation des compétences et la réussite des tests.

Attention cependant : diverses études récentes ont montré que les élèves faisant de l'apprentissage uniquement via des MOOC avaient de moins bons résultats que les élèves passant par des cours en présentiel. L'IA peut-elle arriver à la rescousse ?

Accompagnement d'élèves à déficits cognitifs, qui peut exploiter diverses techniques de captation et de suivi de l'attention[659].

Cobotique qui couvre l'usage de robots accompagnant les enfants dans certains processus d'apprentissage, notamment à destination d'enfants aspergers ou autistes. C'est l'application du petit robot de Leka ainsi que de Nao de Softbank Robotics.

Apprentissage de l'écriture avec l'analyse de l'écriture manuscrite automatique qui fonctionne de manière limitée, et plus efficacement sur la détection de lettres individuelles.

Outils d'accès à la connaissance. L'éducation, surtout dans le secondaire et le supérieur peut bénéficier de l'usage de moteurs de recherche plus intelligents et d'outils voisins de ceux que j'imaginais dans la partie précédente sur les services et le conseil. Ils peuvent être complétés d'autres outils, évoqués dans la rubrique médias et qui permettent de détecter des fausses nouvelles ou informations (« fake news ») aussi bien sur des aspects qualitatifs que quantitatifs.

Accompagnement d'une classe, permettant aux enseignants d'identifier les forces et faiblesses d'un groupe d'élève, les parties bien assimilées ou pas d'un cours, le niveau de leur attention. Des techniques utilisant des caméras et autres capteurs permettent déjà de capter l'attention générale d'une audience et de segmenter les groupes d'individus par comportement.

---

[657] Cet inventaire est inspiré de 7 roles for artificial intelligence in education de Matthew Lynchmay, mai 2018 et de Education and AI from Intelligence Unleashed, An argument for AI in Education de Rose Luckin, UCL Knowledge Lab, University College London.

[658] Voir A college professor used an AI teaching assistant for months, but his students didn't notice, 2016.

[659] Voir How AI is changing special education, juin 2017.

Cela permettrait aussi d'ajuster le contenu de cours car l'attention dépend aussi des pratiques de l'enseignant. Cela fait partie du périmètre de sociétés américaines telles que **Content Technologies** (2005, USA), qui cependant ne cible pas que le secteur de l'éducation, modèle économique oblige[660] et **Carnegie Learning** (1997, USA, acquis par Apollo Education Group, $14M) qui est focalisé sur l'enseignement des mathématiques dans le secondaire et le supérieur[661].

Détection de fraude et de plagiat, une application qui n'est pas souvent évoquée et qui porterait sur les épreuves de concours et sur les thèses, en disposant de versions numériques des textes des élèves, la reconnaissance d'écriture à grande échelle n'étant pas suffisamment fiable.

Aides sur le développement kinétique avec des solutions divers exploitant plus ou moins d'IA dans les sports, la danse, les apprentissages de mouvements dans l'espace en général. L'analyse visuelle des corps permet de diagnostiquer les points d'amélioration. Nous avons aussi vu qu'il existait déjà des générateurs de mouvements synthétiques exploitant des réseaux génératifs. En pratique, pour créer des vidéos avec les visages et le corps des enfants se déplacement sur des mouvements réalisés par d'autres enfants.

Serious games. Dans certains contextes, les jeux peuvent être mis en place pour certains apprentissages, aussi bien de l'écriture que des mathématiques, de physique ou de sciences de la vie. L'IA peut en théorie aider à les conçevoir et à les rendre personnalisables en fonction de chaque enfant. Se posera cependant à chaque fois la question de l'apprentissage de l'IA elle-même et des données qui l'alimentent. S'il s'agit de règles et de faits, on retombe dans des moteurs de règles traditionnels. S'il s'agit d'exploiter des données de jeux d'autres élèves pour faire de l'apprentissage supervisé ou non d'une IA, alors le volume de données nécessaire pourrait dépasser l'entendement. A la fin, l'usage d'IA dans ce créneau dépendra de l'équation économique de la formation. Donc, de son audience. Ce qui favorisera les langues dominantes dans les pays développés, l'anglais en premier.

Analyse de l'attention des élèves et de la performance des enseignants. Là encore, des outils vidéo de détection des émotions et de l'attention pourraient permettre de suivre les élèves inattentifs ainsi qu'évaluer les enseignants qui n'arrivent pas à capter l'attention de leurs élèves. Ce sont cependant des domaines où il faut se méfier du solutionnisme technologique selon lequel on peut résoudre tous les problèmes avec de la technologie. Ici, la méthode est contestable et probablement difficile à justifier économiquement pour l'équipement de toutes les salles de classes. Mais le monde est grand et cela sera sûrement expérimenté quelque part si ce n'est pas déjà fait.

Education à l'IA. Tous les outils cités précédemment sont de l'IA appliquée à l'éducation. Reste à s'intéresser à l'enseignement de l'IA elle-même. Il est probable qu'il se positionnera dans la lignée des enseignements du numérique avec d'un côté des enseignements sur les usages des outils exploitant de l'IA, et de l'autre, à la création de ces outils pour les futurs professionnels du secteur. Cela passe par l'explication des concepts de base de l'IA, la description des applications dans le raisonnement automatique, le traitement des données, celui des langages et de la vision, puis des systèmes qui les embarquent (mobiles, robots, véhicules autonomes, chatbots, agents vocaux…). Pour les techniciens et ingénieurs, il faudra passer par des mathématiques, des concepts nombreux autour des réseaux de neurones, des outils de développement, de la data science, de l'architecture de systèmes et de l'intégration.

Recrutement. L'IA est mise en œuvre aux USA dans les universités pour les aider à recruter les meilleurs étudiants comme chez **Plexuss** (2014, USA, $3,1M) ou, au contraire, pour aider ces derniers à trouver la meilleure université avec **Admitster** (2014, USA).

---

[660] Content Technologies cible aussi les marchés de la santé, du secteur public et de la finance.

[661] Citées dans How Is AI Used In Education - Real World Examples Of Today And A Peek Into The Future, de Bernard Marr en juillet 2018.

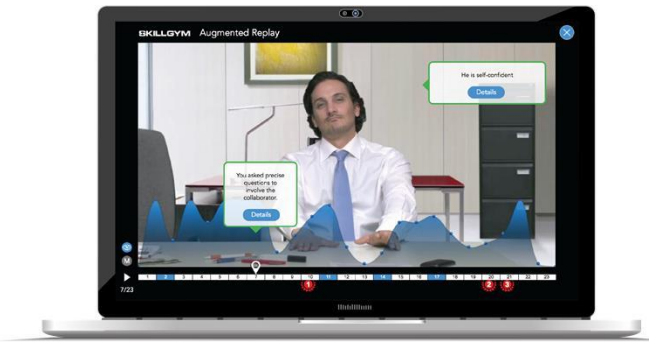
Automatisation des processus administratifs. Et là, il y a du boulot ! La plupart des établissements sont loins d'avoir fait leur « transformation digitale ». J'en fais moi-même l'expérience chaque année lorsque je dois remplir n fois un dossier de vacataire pour mes conférences dans l'enseignement supérieur. Les formulaires sont remplis sur papier. La modernitéconsiste à annoter un PDF qui peut être envoyé par email. Il n'y a pas de formulaire en ligne. Il faut remplir le même formulaire chaque année et fournir à chaque fois les mêmes pièces administratives. Bref, c'est l'âge de la pierre. Et vous croyez que c'est là que l'on va développer des solutions d'IA ? Je crains fort que cela ne soit pas le cas, tout du moins en France. Tout du moins pour les processus concernant les enseignants. Pour les élèves, pourquoi pas, là où ils sont nombreux et dans les grands établissements et grandes filières. Au passage, la majorité des processus administratifs ont besoin d'applications numériques traditionnelles et l'IA est superfétatoire dans un premier temps.

Startups

J'ai identifié quelques startups dans le monde qui s'attaquent aux usages de l'IA dans l'éducation et de manière très différente. Comme partout dans ce document, cet inventaire est bien plus illustratif qu'exhaustif.

- **AiEducate** (UK) a créé DALI (Dialog-based Artificial Learning Intelligence), un chatbot de tutorat. Le chabot joue deux rôles : il gère des conversations en mode questions/réponses classiques et enseigne. Ils sont partenaires d'IBM Watson, ayant participé au concours Xprize d'IA organisé par ce dernier. Il n'est pas évident que ce projet ait une vie commerciale.
- **AskMyClass** (USA) s'intègre avec Amazon Alexa et Google Home, pour faire faire des exercices aux élèves du primaire, dans l'apprentissage du vocabulaire. Il permet aussi aux enseignants de prendre des notes. Le chatbot est censé faire gagner plus d'une heure de travail aux enseignants par semaine. La solution est commercialisée en mode freemium. C'est évidemment adapté au marché US ou tout du moins anglophone.
- **Gradescope** (2014, USA, $5,3M) réduit par deux le temps que les enseignants passent pour noter les copies des élèves d'examens passés en ligne. Cela passe par de la reconnaissance d'images et fonctionne visiblement pour les sciences « dures » (maths, physique, chimie, biologie, informatique). Le système regroupe les réponses par catégories ce qui permet de noter une seule fois chaque réponse différente.
- **Skillogs** (2015, France) a développé Acarya, du Sanskrit "Professeur", un environnement d'apprentissage qui permet un apprentissage adapté à la vitesse de chaque élève, selon leur niveau (vidéo). L'IA utilisée ? Bien, elle n'est pas précisée, comme c'est souvent le cas. On ne sait en particulier pas identifier les données d'entraînement de cette IA. Il s'agit d'une banque de données de compétences, d'analyse des forces et les faiblesses des élèves. C'est encore une solution qui s'appuie sur IBM Watson.
- **Domoscio** (2013, France, $569K) propose un « *outil d'ancrage adaptatif mémoriel* » fondé sur l'IA et le machine learning (pourquoi cette distinction ? L'IA pour le traitement du langage et le ML pour l'analyse des données ?). Il analyse le niveau et la manière d'apprendre des élèves pour leur proposer le contenu adapté et au bon moment, histoire de favoriser la mémorisation. C'est visiblement plutôt utilisé pour la formation continue, dans de grandes entreprises (SNCF, Banque de France, Bouygues Télécom) et dans l'enseignement supérieur (Sciences Po, Universités).
- **Soul Machines** (Nouvelle-Zélande, $7,5M) a développé des avatars très réalistes, qui sont intégrés dans la solution de formation sur les énergies renouvelables de **Vector** (vidéo). On n'est pas très loin du syndrome de la vallée de l'étrange, ce sentiment curieux et dérangeant que l'on ressent (en Occident, pas en Asie) lorsque l'on interagit avec une machine qui adopte l'apparence humaine.

- **Groupe Bizness** (2007, France) utilise #skillgym pour la formation sur-mesure dans la vente, le management, la transformation digitale (?), les relations et la satisfaction client. Ils testent la réaction émotionnelle des élèves dans des mises en situation (via une tablette et leur webcam) avec des vidéos interactives jouées par des acteurs. Leurs clients sont dans la banque, les assurances et l'automobile. (vidéo)[662].



- **Snapask** (2015, Hong Kong, $21,8M) a développé une application qui permet de mettre en relation des élèves et des tuteurs pour des séances de coaching thématiques. Les élèves prennent une photo du problème à résoudre et le décrivent avec quelques mots. L'application se charge ensuite de la mise en relation avec le tuteur le plus approprié. Ils affichent en avoir déjà 120 000. Ils sont rémunérés à hauteur de $1 par question traitée. La startup n'est pas encore véritablement internationalisée. C'est une idée parmi d'autres.

Services publics

Au même titre que le numérique de manière générale, l'IA peut être mise en œuvre quasiment partout dans les processus des services publics. Dès lors qu'il y a des processus et des données ou de l'information, l'IA peut jouer un rôle. Le cas le plus classique est l'amélioration de la relation avec les citoyens ou clients[663].

Cela peut notamment passer par l'usage de chatbots. Les données chiffrées de nombreuses activités peuvent servir à réaliser des prévisions, comme pour faire des simulations de rentrées ou de dépenses fiscales.

Les applications sont plus nombreuses dans le contexte de la ville intelligente pour prédire l'usage des infrastructures, optimiser la consommation d'énergie ou gérer la sécurité par la vidéo-surveillance.

L'IA peut aussi servir pour la police et la justice. La police de la ville de **Durham**, au Royaume-Uni expérimentait en 2017 une application développée par l'Université de Cambridge servant à classifier les suspects arrêtés pour évaluer leur niveau de risque, exploitant quatre années d'archives d'arrestations. Le système est dénommé HART pour **Harm Assessment Risk Tool**[664]. La startup **Predpol** (2012, USA, $3,7M) a déjà une offre avec ce même positionnement pour prédire les lieux et les fenêtres temporelles où des crimes pourraient être commis.

---

[662] L'usage de l'IA dans la formation continue fait écho à l'impact de l'IA sur les métiers. The role of education in AI (and vice versa), avril 2018, évoque les compétences à acquérir dans un monde professionnel entouré d'IA. Le gagnant est le développement de la créativit et des compétences sociales et de perception. Bref, des *soft skills*.

[663] Voir Artificial Intelligence for Citizen Services and Government, août 2017 (19 pages).

[664] Le digital evangelist Stéphane Mallard dans ses conférences, comme dans L'intelligence Artificielle - A l'aube de la disruption ultime, indique que ce système permet de prévoir les crimes à l'avance, avec la date et lieu. Ce n'est pas du tout la fonction de HART ! Comme de nombreux évangélistes du secteur, les exemples donnés qui relèvent d'une revue de presse de premier niveau sont souvent très exagérés dans leur portée et leur fonction réelle. En pratique, et pour ce qui concerne les anglais, cette fonctionnalité est anticipée pour 2030. Aujourd'hui et 2030, ce n'est pas la même chose ! Voir The real Minority Report: By 2030, police could use AI to predict and prevent crimes BEFORE they happen, septembre 2016. Les exagérations de ce genre sur l'IA sont très courantes.

Avec le risque de générer des biais liés aux données d'entraînement, focalisant les forces de l'ordre sur les personnes de couleur, dénoncé dans plusieurs études[665]. Une expérience similaire a été menée en Inde avec la startup israélienne **Cortica**[666] qui exploite la vidéo-surveillance. Il en va de même pour l'expérience menée par **Palantir** à New Orleans sans prévenir la ville[667] !

UVeye (2016, Israël, $4,5M) utilise le traitement de l'image de manière originale, en scannant les dessous des véhicules pour détecter des menaces comme des bombes, armes ou caches diverses. Le système s'appuie sur une vision en 3D. Reste à récupérer de bonnes données d'entraînement (vidéo).

Evolv Technology (2013, USA, $30M) est un scanner corporel qui détecte les armes et explosifs avec moins de contraintes que les habituels scanners d'aéroports, à savoir que l'on n'est pas obligé de mettre de côté les objets métalliques. Le débit est de 600 personnes par heures ce qui reste encore faible (un toutes les six secondes). Le système utilise aussi la reconnaissance faciale pour détecter des personnes recherchées (vidéo).

Deep Sentinel (USA, $7,4M) propose un système de caméra de surveillance domestique qui détecte les crimes avant qu'ils interviennent. Le système détecte toute activité anormale par rapport aux habitudes et déclenche l'alarme si besoin est. De là à en faire un PreCog de Minority Report, il ne faudrait pas exagérer !

La **Chine** expérimente un système équivalent qui ambitionne d'aller plus loin en tentant de prévoir où des crimes pourraient avoir lieu en suivant les déplacements de groupes de criminels connus[668]. Un tel système ne peut fonctionner correctement que s'il dispose d'une base de donnée de ces suspects et s'il est capable de suivre leurs déplacements en temps réel. L'exploitation d'images de caméras de vidéo-surveillance peut servir à cela et à détecter des comportements suspects comme ceux des pick-pockets[669].

Dans le même domaine, citons enfin cette expérience menée en Hongrie de détecteur de mensonge à base d'IA utilisé dans le contrôle des frontières pour détecter l'immigration illégale[670]. Le système utilise un agent virtuel qui pose des questions. Le visage des répondants est analysé par l'IA avec des techniques probablement voisines de celles d'Affectiva. Les premiers résultats étaient exacts à 75%, ce qui est bien faible. L'équipe du projet **iBorderCtrl** pense que ce taux pourra atteindre 85%.

Renseignement et défense

L'IA a évidemment de nombreuses applications dans le vaste secteur de la défense et du renseignement[671]. Ce sont d'ailleurs de gros financeurs des recherches dans ce champ scientifique et technologique, notamment via l'agence DARPA du Pentagone, qui finance des concours, challenges et appels à projets de R&D appliquée. Ceci active de nombreux laboratoires d'universités, startups et entreprises établies. En juin 2018, le Département de la Défense US lançait le JAIC (Joint Artificial Intelligence Center) pour piloter la recherche appliquée sur l'IA avec un budget de $1,7B étalé sur cinq ans.

---

[665] Voir Pitfalls of Predictive Policing, septembre 2016, qui fait reference à l'étude de la Rand Corporation sur l'expérience menée à Chicago ici A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot.

[666] Voir Crime-predicting A.I. isn't science fiction. It's about to roll out in India, avril 2018.

[667] Voir Palantir has secretely been using New Orleans to test its predictive policing technology, février 2018.

[668] Voir China seeks glimpse of citizens' future with crime-predicting AI en juillet 2017.

[669] Encore faut-il alors intervenir ! On n'a pas besoin d'IA pour identifier les pick-pockets dans les lieux touristiques à Paris et ils ne sont pas arrêtés pour autant. Voir Chinese Police Arrest Suspect At Pop Concert Using Facial Recognition, mars 2018.

[670] Voir An AI Lie Detector Is Going to Start Questioning Travelers in the EU, de Melanie Ehrenkranz, octobre 2018. On plaint les personnes qui seront victims de faux positifs de ce genre de système et seront bloqués longtemps au contrôle des frontières.

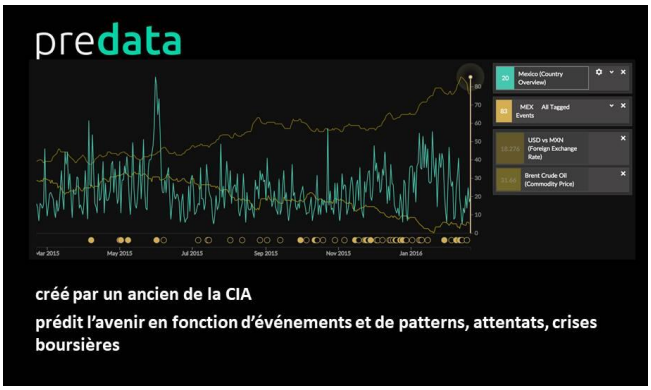
[671] Voir Artificial Intelligence and National Security de Greg Allen et Taniel Chan, 2017 (132 pages) et Artificial Intelligence and National Security de Daniel Hoadley et Nathan Lucas, 2018 (42 pages).

Ce JAIC dépend du DSI du DoD et pas de la DARPA[672].

La robotique est déjà très largement utilisée, que ce soit avec des **robots de déminage** déjà opérationnels en Irak ou en Afghanistan (mais ils sont télécommandés) et surtout avec les **drones aériens**, eux-aussi télécommandés mais dotés d'outils de reconnaissance de leur environnement et de pilotage automatique. Ce sont soit des drones de reconnaissance, soit des drones lanceurs de missiles. Les robots tueurs entièrement autonomes ne sont pas encore opérationnels, mais rien n'empêchera les pays en roue libre d'en créer, malheureusement, même si comme le chercheur anglais Stuart Russell en fait la promotion, l'ONU déclare ces robots « illégaux »[673].

L'IA est aussi utilisée dans la prévision d'événements et de risques géopolitiques comme chez **Predata** (2015, USA, $3,25M).

C'est un bon exemple d'usage sophistiqué de l'IA pour un exécutif, et qui pourrait aussi s'appliquer aux prévisions dans la création de politiques économiques. Mais celles-ci sont toujours créées en fonction d'un croisement entre idéologies partisanes, méthode Coué et en jouant avec les contraintes de l'administration.



Dans le renseignement et l'analyse des interceptions de télécommunications, l'IA est évidemment utilisée, notamment avec des outils de traitement du langage. D'autres outils, basés sur du machine learning, permettent de détecter des bizarreries comportementales. Cela peut faire appel à du machine learning non supervisé servant à identifier des clusters atypiques de comportements. Les comportements peuvent concerner les déplacements ou les communications de personnes ou de groupes de personnes surveillées.

Evidemment, ce n'est pas très bien documenté avec de belles études de cas de startups. La **NSA** comme la **DGSE** ou le **GCHQ** de nos voisins anglais ne sont pas du tout bavards sur la question.

L'IA peut ainsi servir à identifier des terroristes potentiels en fonction de profiling exploitant diverses sources de données, issues notamment de l'espionage des communications électroniques. Ces outils utilisent du machine learning, de la PCA (Principal Components Analysis) pour identifier les paramètres permettant de les distinguer du reste de la population. L'IA sert à trouver des aiguilles dans de vastes bottes de foin.

Digital Reasoning (2000, $73M) a été créée par des anciens d'Oracle et de la CIA (entre autres provenances) et est financée par In-Q-Tel, le fonds d'investissement de cette dernière. Sa solution d'analyse de données est utilisée par le renseignement et la défense US ainsi que dans la finance. Comme celle de Skymind, sa solution Synthesys est en Java et ouverte. Elle permet d'analyser des données structurées et non structurées, y compris des conversations téléphoniques. Elle sert à détecter des comportements anormaux dans les communications électroniques. C'est donc un outil utilisé par la NSA dans la gestion de ses interceptions (PRISM & co).

L'Université de Caroline du Sud a même développé une solution qui prédit l'émergence de protestations violentes en surveillant les réseaux sociaux, un projet financé par le Ministère de la Défense[674]. Dans un pareil cas, on s'approche des caractéristiques d'un état policier, réprimant toute forme de protestation comme c'est le cas en Russie. L'Internet est aussi utilisé comme un outil de répression et de régression des libertés dans les démocraties !

---

[672] Voir Establishment of the Join Artificial Intelligence Center, juin 2018 (2 pages).

[673] Voir Risques des robots tueurs, armes autonomes et les solutions pour y faire face de Dimitri Carbonnelle, mai 2018.

[674] Voir Social media posts may signal whether a protest will become violent, mai 2018.

Acteurs de l'IA

Une fois que l'on a fait le tour des algorithmes et technologies de l'IA et de ses usages potentiels souvent issus de nombreux laboratoires de recherche, il faut passer par la case « action » et avancer.

Pour cela, les entreprises ont besoin de faire appel à un panaché d'acteurs qu'il nous faut examiner : les grandes entreprises du numérique dont les GAFAMI, les startups et les entreprises de services du numérique (ESN).

Nous allons ici examiner ces catégories d'acteurs, leurs offres, forces et faiblesses. Nous terminerons avec un tour de l'écosystème français de l'IA, en y intégrant la recherche.

Grandes entreprises du numérique

Les grands acteurs du numérique occidental sont tous très impliqués dans l'IA pour améliorer leurs solutions. Nous avons en tête les GAFAMI : Google, Amazon, Facebook, Apple, Microsoft et IBM. Puis d'autres grands éditeurs de logiciels comme Oracle, SAP et Salesforce.

Tous ne jouent pas le même rôle dans les grandes entreprises. Seuls IBM, Microsoft, Amazon et dans une seconde mesure Facebook et Google, proposent des plateformes et solutions adaptées aux grandes entreprises.

IBM et Microsoft sont les entreprises investies dans l'IA depuis le plus longtemps. L'un des initiateurs du Summer Camp de Darmouth en 1955 était Nathaniel Rochester, un chercheur d'IBM. Et Microsoft a créé son laboratoire de recherche en 1991, principalement dédié aux avancées de l'IA, traversant des époques où cela n'était pas du tout à la mode.

Les GAFAMI ont la particularité de publier en open source presque toutes leurs briques logicielles de base de l'IA. Ce sont donc des commodités. Vu de loin, il n'est pas évident de départager les outils de reconnaissance d'image et du langage que l'on peut trouver chez IBM, Google, Amazon et Microsoft. Les différences se situent plus dans les modèles économiques et leur manière d'aborder le marché des entreprises.

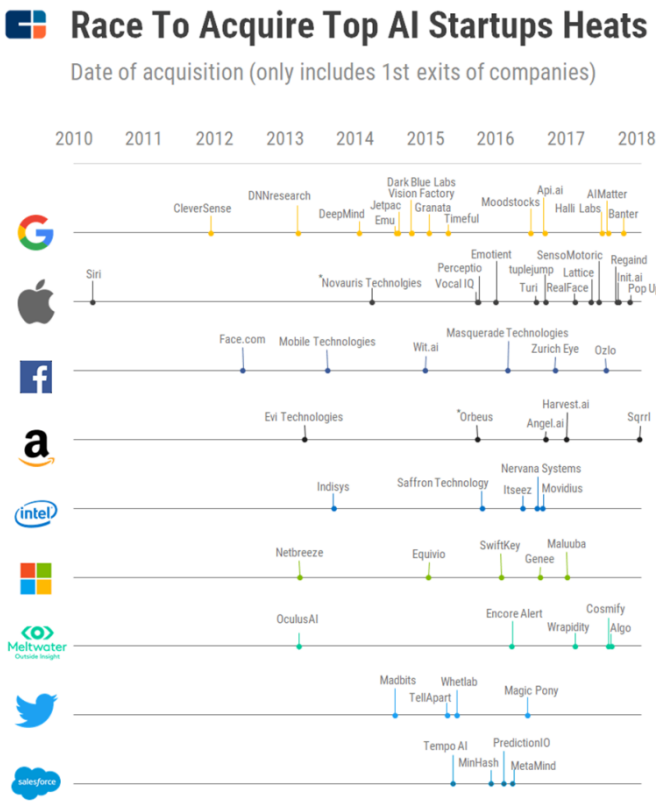
La réussite dans l'IA comprend quelques ingrédients de base : des plateformes logicielles intégrées et ouvertes, des ressources en cloud éventuellement enrichies par des architectures matérielles propriétaires différenciées, l'accès à des données sachant que les plus intéressantes et volumineuses proviennent d'activités grand public, des partenaires adoptant la plateforme, ou la capacité interne à gérer des projets clients et enfin, surtout, des talents, qui sont de plus en plus difficiles à dénicher.

Sur les grands marchés horizontaux, ces GAFAMI et leurs équivalents chinois sont déjà les leaders mondiaux de l'IA[675]. Il reste cependant encore des places à prendre pour des acteurs positionnés sur des marchés verticaux que ces grands acteurs ciblent mal.

J'évoquerai le cas des grands acteurs du numérique chinois à part dans la rubrique sur la géopolitique de l'IA. Ceux des acteurs à observer de plus près au sujet de l'IA sont Huawei et Baidu. Le premier avec sa stratégie d'IA « end to end » allant des infrastructures d'opérateurs télécoms jusqu'aux chipsets de smartphones (Kirin 970/980) et le second avec sa stratégie de plateforme DuerOS et surtout Apollo pour les véhicules autonomes.

[675] Voir Why AI consolidation will create the worst monopoly in US history, août 2016.

L'autre benchmark habituel des grands groupes pure players du numérique consiste à observer leurs acquisitions. Comme il y a une guerre de talents dans l'IA, c'est un indicateur de leur montée en puissance. De ce côté-là, Google est plutôt en tête de peloton suivi par Apple qui est devenu un plus gros acquéreur de startups depuis quelques années. Le schéma ci-contre est issu de CBInsights en février 2018. Ces acquisitions sont soit technologiques, soit de l'ordre de l'acquihire, un barbarisme décrivant le recrutement de talents au prix fort. Il faut aussi intégrer de nombreux investissements « préventifs ».



Les GAFAMI adoptent aussi une double approche de **plateforme** pour attirer des développeurs d'applications avec des frameworks le plus souvent open source et générer des économies d'échelle et d'**intégration verticale** pour capter une partie aussi grande que possible de la valeur ajoutée, qu'il s'agisse de sa dimension technique ou des usages, parfois jusqu'à cibler des marchés verticaux.

Cette intégration verticale couvre de nombreux domaines comme les chipsets dédiés à l'IA sur serveurs ou embarquée (IBM, Google, Microsoft, Intel, Apple), les applications et données grand public (tous sauf IBM), les données métier (surtout IBM), la mobilité (Google, Apple, Facebook), la réalité virtuelle (Google, Facebook, Microsoft), la création et/ou la distribution de contenus (Google, Amazon, Apple, Facebook) et plus rarement, les services, le conseil et l'intégration (surtout chez IBM et Oracle).

Voici un petit résumé comparatif de la situation des GAFAMI et autres grands acteurs du numérique :

- **IBM** est un acteur très présent dans les grandes entreprises avec sa plateforme logicielle « couteau suisse » IBM Watson et son activité de services associé, sans compter ses offres verticales comme dans la santé ou la finance. Et la société est investie dans l'IA depuis ses débuts, en 1955 avec Nathaniel Rochester.
- **Google** est un acteur dominant du numérique grand public couvrant de larges pans de la vie des utilisateurs (recherche, email, mobile, TV, maison connectée). Cela lui permet de capter d'énormes volumes de données pour alimenter « ses IA ». Il domine aussi le développement de solutions d'IA avec son framework open source TensorFlow. C'est un opérateur de cloud qui a l'ambition de servir aussi les entreprises.
- **Amazon** est le leader de la distribution en ligne et du cloud d'entreprise. Il s'est taillé une bonne place dans les assistants vocaux avec Alexa, qui est maintenant très sérieusement concurrencé par Google Assistant. Sa plateforme logicielle semble moins influente que celles de Google et IBM.
- **Microsoft** propose une plateforme logicielle complète d'IA couvrant le machine learning et le deep learning, sur les grands pôles que sont le traitement des données, du langage et de l'image, ainsi qu'une excellente activité de recherche fondamentale. C'est aussi un important opérateur de cloud, derrière Amazon. L'éditeur souffre cependant d'un déficit marketing dans le domaine, notamment auprès des startups qui sont peu nombreuses à adopter ses SDK et APIs.

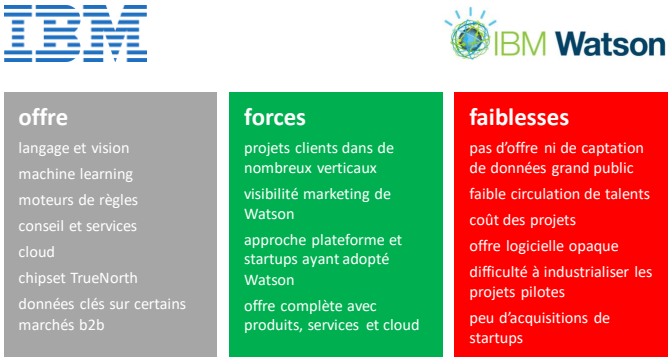
- **Facebook** domine les réseaux sociaux et la communication mobile, des logiciels qui exploitent de plus en plus d'IA, aussi bien dans le traitement du langage (Messenger) que de l'image (WhatsApp et Instagram). Par contre, à part ses APIs pour créer des chatbots pour Facebook Messenger et le framework PyTorch, Facebook est modérément influent auprès des développeurs. Tout du moins face à Google, IBM et Microsoft.
- **Apple** est une société très intégrée verticalement dont l'approche plateforme, surtout en cloud, est moins prégnante dans l'industrie. iOS étant un passage obligé, il est mécaniquement présent chez de nombreux développeurs. Apple ne joue pas un grand rôle dans l'IA des entreprises.
- **Intel** et **Nvidia** sont les deux leaders des processeurs pour serveurs et dans une certaine mesure dans l'embarqué, même si ce dernier marché est très fragmenté avec de nombreux acteurs. Intel a une offre comprenant une myriade de CPU et processeurs neuromorphiques pour serveurs et l'embarqué, certains étant adaptés au traitement du langage et d'autres, de l'image. Nvidia a une offre assez cohérente de GPU avec des chipsets intégrant des unités de traitement classiques (ALU) et des tenseurs (multiplicateurs de matrices) qui sont adaptés au machine learning et au deep learning.
- **Oracle**, **SAP** et **Salesforce** ont mis l'IA à leur menu mais font moins parler d'eux du fait de leur positionnement entreprise. Salesforce est probablement le plus avancé des trois dans les usages de l'IA.

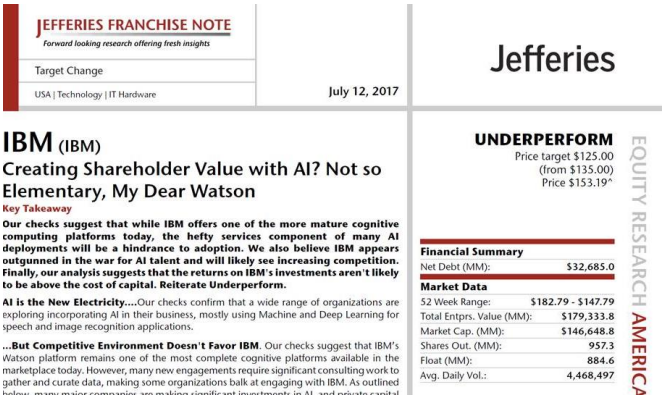
Toutes choses que nous allons examiner un peu plus en détail dans ce qui suit, acteur par acteur !

IBM

IBM est l'un des premiers grands acteurs du numérique qui ait mis le paquet sur l'IA relativement tôt, au moins à partir de 1996. IBM articule l'IA autour du concept d'informatique cognitive et de la marque un peu fourre tout IBM Watson.

Une analyse au vitriol de 53 pages publiée en juillet 2017 par la banque d'investissement Jefferies[676] décrit bien les enjeux d'IBM vus par la lorgnette des actionnaires : Watson est un magnifique artifice de communication, mais sa traduction en avantage compétitif n'est pas évidente pour l'entreprise dont la stratégie est tiraillée entre celle de prestataire de service et d'éditeur de logiciel. Et elle fait face à des concurrences multiformes.




JEFFERIES FRANCHISE NOTE
Forward looking research offering fresh insights
Target Change
USA | Technology | IT Hardware
July 12, 2017
Jefferies
IBM (IBM)
Creating Shareholder Value with AI? Not so Elementary, My Dear Watson
Key Takeaway
Our checks suggest that while IBM offers one of the more mature cognitive computing platforms today, the hefty services component of many AI deployments will be a hindrance to adoption. We also believe IBM appears outgunned in the war for AI talent and will likely see increasing competition. Finally, our analysis suggests that the returns on IBM's investments aren't likely to be above the cost of capital. Reiterate Underperform.
AI is the New Electricity....Our checks confirm that a wide range of organizations are exploring incorporating AI in their business, mostly using Machine and Deep Learning for speech and image recognition applications.
...But Competitive Environment Doesn't Favor IBM. Our checks suggest that IBM's Watson platform remains one of the most complete cognitive platforms available in the marketplace today. However, many new engagements require significant consulting work to gather and curate data, making some organizations balk at engaging with IBM. As outlined
UNDERPERFORM
Price target $125.00
(from $135.00)
Price $153.19*
Financial Summary
Net Debt (MM): $32,685.0
Market Data
52 Week Range: $182.79 - $147.79
Total Entprs. Value (MM): $179,333.8
Market Cap. (MM): $146,648.8
Shares Out. (MM): 957.3
Float (MM): 884.6
Avg. Daily Vol.: 4,468,497
EQUITY RESEARCH AMERICA

Elle comprend les grandes sociétés de services et intégrateurs (CapGemini, Atos, Orange Business Services pour la France), les grands éditeurs de plateformes (Google, Microsoft, Oracle, …) et les fournisseurs de cloud (Amazon et Microsoft). IBM a aussi bien du mal à attirer les talents qui sont aspirés par les GAFA et les startups de la Silicon Valley. Au point que début septembre 2017, IBM annonçait le lancement d'un laboratoire conjoint de recherche avec le MIT financé à hauteur de $24M par an sur 10 ans.

[676] Voir IBM Creating Shareholder Value with AI? Not so Elementary, My Dear Watson, Jefferies Franchise Note, juillet 2017.

L'histoire

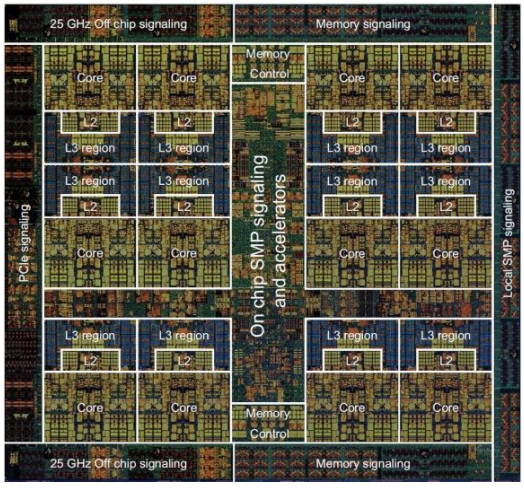
Dans les années 1960, IBM aurait stoppé brutalement ses travaux de recherche en IA par peur que les postes de managers soient remplacés par des machines. C'était aussi le résultat d'une remontée des clients dans les DSI qui avaient aussi peur de perdre leur poste. L'histoire a tendance à se répéter pour ce qui est des réactions du marché. On sous estime d'ailleurs les effets de boucle rétroactive entre les promesses effrayantes que certains relaient sur l'IA et la réaction de rejet ou de prudence que cela peut générer dans les entreprises.

IBM a depuis fait sa mue de constructeur vers le métier d'éditeur de logiciels couplé à celui de prestataire de services à partir de 1993. Aujourd'hui, IBM est une société à nouveau en lent déclin, en tout cas en termes de chiffres d'affaires.

IBM génère maintenant l'essentiel de son profit à parts égales entre logiciels et services. La synergie entre les deux métiers est plutôt bonne même si la branche services d'IBM travaille aussi avec les technologies concurrentes. Ils savent déployer des solutions qui intègrent des logiciels d'Oracle, de Microsoft, de SAP, bref de tout, en fonction des contraintes du client.

La question reste cependant pour tout acteur du marché de ne pas rater les vagues technologiques. IBM s'en était pas trop mal sorti en 2000 en se positionnant dans le e-business. Sa campagne de communication martelait le rôle de fournisseur "one-stop-shop" pour ses clients.

IBM a petit à petit délaissé ses activités matérielles dans les machines de commodité. Le délestage s'est fait par étapes : les imprimantes avec la création de Lexmark en 1991, les PC cédés en 2004 au chinois Lenovo, et puis les serveurs PC cédés également à Lenovo, en 2014. Par contre, ils ont toujours misé sur les grandes architectures, dans la lignée de leur ligne historique de mainframes. D'où l'importance pour eux du HPC (High Performance Computing) et de l'intelligence artificielle. Il leur reste les mainframes de la série Z et les serveurs Power9 équipés du chipset du même nom (*ci-contre*), avec 12 cœurs et produit en 14 nm FinFET sur wafers SOI (silion on insularor) d'origine SOITEC.



Ils investissent aussi dans des chipsets neuromorphiques et dans l'informatique quantique. La recherche fondamentale d'IBM reste l'une des plus actives dans le secteur privé.

La première incartade d'IBM dans l'IA s'est manifestée au grand jour avec la victoire de l'ordinateur **IBM Deep Blue** (appelé initialement Deeper Blue) contre Gary Kasparov en 1997. Cela a contribué à relancer les recherches d'IBM sur l'IA dans les années 2000.

La seconde grande étape a été la victoire d'IBM Watson au jeu **Jeopardy** en 2011. Jeopardy est une sorte de "Questions pour un Champion" américain, sans Julien Lepers. Watson n'est pas infaillible. Cette victoire fut un peu enjolivée et construite par la communication d'IBM qui au passage, a été pilotée à l'échelle mondiale par l'agence **Ogilvy**[677].

[677] Dans une partie intéressante et moins médiatisée (vidéo) organisée avec Miles O'Brien et David Gondek, l'un des créateurs de Watson, Watson ne sait pas indiquer pendant quelle décennie Klaus Barbie a été condamné ni indiquer sur quelle place de Dallas (Dealey Plaza) JFK a été assassiné, ni ce qu'est la vermiphobia (la phobie des vers) ou la ailuraphobia (phobie des chats), toutes ces informations étant pourtant disponibles sur Wikipedia. Il ne savait pas non plus identifier des recettes de cuisine en fonction de leurs composantes. Watson avait aussi du mal à répondre à des questions formulées avec peu de mots et comprenant des ambiguïtés ou des doubles sens. Tout est question de base de connaissances. Celle-ci comprenait 200 millions de pages de données structurées et non structurées représentant un total de 4 To, toutes chargées en mémoire pour assurer un temps de réponse rapide. C'était d'ailleurs bien injuste, car les joueurs humains n'avaient pas le droit d'accéder à Wikipedia sur leur mobile. A l'époque, les smartphones étaient déjà de la partie ! La mémoire verbale humaine est à peine de quelques Go. Elle ne peut pas concurrencer une mémoire de 4 To !

Watson était au départ un projet de recherche baptisé BlueJay (2007) focalisé sur l'exploitation de gros volumes de données non structurées. Il s'intégrait dans la volonté d'IBM Research de s'attaquer à un grand défi, la réussite du fameux test de Turing. Watson était d'abord présenté comme un ordinateur, s'appuyant sur une architecture massivement parallèle à base 750 serveurs utilisant des processeurs Power7 octo-cœurs tournant à 3,5 GHz et totalisant 16 To de RAM.

IBM Watson est devenu une plate-forme logicielle, respectant en cela les canons de la réussite dans le numérique[678]. Elle est proposée aux développeurs sous forme d'APIs (interfaces de programmation) dans le cloud. Watson s'appuie principalement sur la solution **DeepQA** d'IBM et le framework **Apache UIMA** (Unstructured Information Management Architecture) qui permet d'exploiter des données non structurées et de créer des chatbots. Suivent d'autres briques de traitement du langage et pour la vision artificielle.

En 2011, Watson devenait le sujet phare de la communication d'IBM autour de la notion de « cognitive computing », une appellation qui leur a permis d'éviter aux débuts celle de l'IA qui n'avait pas encore bonne presse. IBM englobe dans cette appellation presque toutes les applications qui gèrent… de l'information, qu'elles s'appuient ou pas sur des briques logicielles d'IA, y compris la gestion de bases de données transactionnelles.

Depuis 2016, IBM utilise les deux. L'IA était le seul thème de l'intervention en **keynote** de Ginni Rometty au CES de Las Vegas de janvier 2016 et elle va remettre cela en étant la principale intervenante en keynote au CES 2019 en janvier. IBM organise aussi des conférences autour de Watson, d'abord "World of Watson" puis « Think »[679].

En France, IBM se focalise notamment sur le développement des compétences autour de l'IA. Ils sont pour ce faire partenaires de l'**emlyon business school** avec qui ils développent un « GPS des compétences », une plateforme d'anticipation des tendances en matière d'emplois, de modèles de formations et d'accompagnement, avec l'**Ecole Nationale Supérieure de Cognitique** de Bordeaux, pour créer une chaire en Sciences Cognitives avec une formation en IA, avec l'**ESIEE Paris**, pour un autre chaire « IA et prescriptive analytics » et avec **Grenoble INP-ENSIMAG**, pour l'accès à des ressources de calcul en cloud. Tout cela, bien entendu, autour d'IBM Watson.

Les logiciels

A chaque solution d'IA, son assemblage de composants hétéroclites réalisé sur mesure pour répondre à un besoin. C'est particulièrement vrai d'IBM Watson. Ce dernier est un très bon coup business et marketing d'IBM, qui a réussi à simplifier partiellement un sujet très complexe.

Ils ont ainsi vulgarisé les capacités de Watson et pu cacher sa complexité, voisine de celle de l'architecture de WebSphere. IBM Watson est comme le fakir du célèbre sketch de Pierre Dac et Francis Blanche[680] : dès que l'IA peut jouer un rôle dans un projet, « il peut le faire ».

Mais IBM Watson n'est pas ni produit ni une solution ni « une IA ». C'est une architecture et une plateforme faite de nombreuses briques logicielles qu'il faut assembler et au-dessus de laquelle il faut développer des solutions logicielles sur mesure pour en tirer quoi que ce soit.

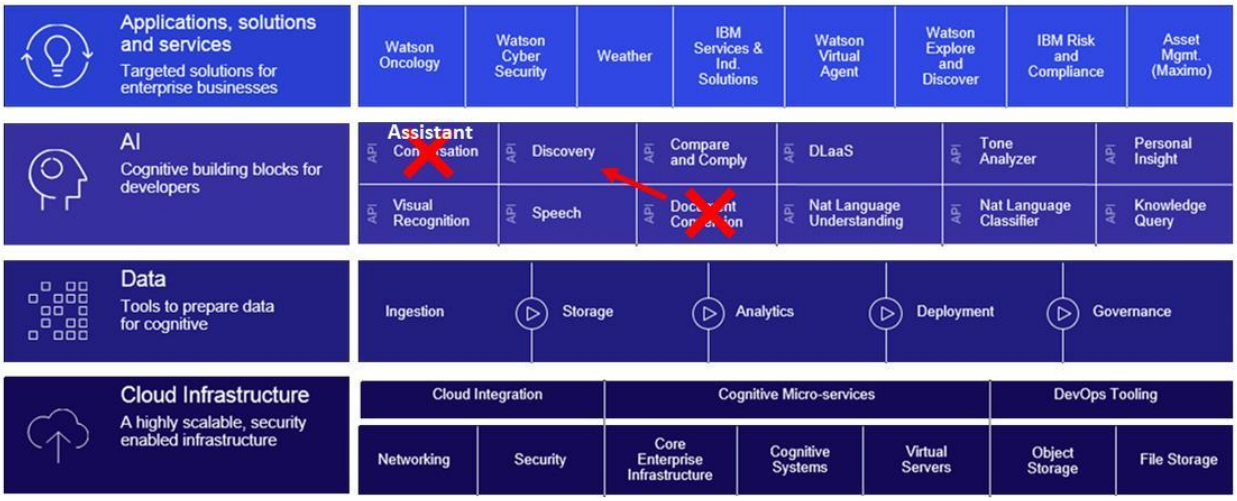
IBM passe son temps à changer les noms de ses briques d'une année à l'autr !

---

[678] L'histoire est bien racontée dans IBM Watson: The inside story of how the Jeopardy-winning supercomputer was born, and what it wants to do next de Jo Best, 2015.

[679] Les dernières conférences IBM Think avaient lieu en mars 2018 à Las Vegas, octobre 2018 à Paris et la suivante est prévue en février 2019 à San Francisco.

[680] Voir Le Sar Rabindranath Duval de 1956 à 7 minutes 44s qui comprend d'ailleurs le sketch « Le biglotron » qui constitue une excellente description prémonitoire d'IBM Watson dans à 9 minutes et 50 secondes qui date de 1958. Et une variante avec La voyante Arnica qui date de 1957, à partir de la cinquième minute.



Je n'ai pas réussi à trouver une version publique plus récente du schéma *ci-dessus* alors que Conversation devenait Assistant, que Document Conversion était intégré dans Discovery, sans compter le fait que le module Machine Learning ne figure pas dans ce schéma ni leur moteur de règles issu du rachat d'ILOG il y a 10 ans. Les modules « cognitifs » sont à l'exception de celui de la vision tous liés au traitemet du langage.

IBM Watson est proposé aux développeurs de solutions sous la forme d'APIs REST[681] qui permettent d'accéder à une large panoplie de services, qui sont intégrées dans la plateforme en cloud Bluemix.

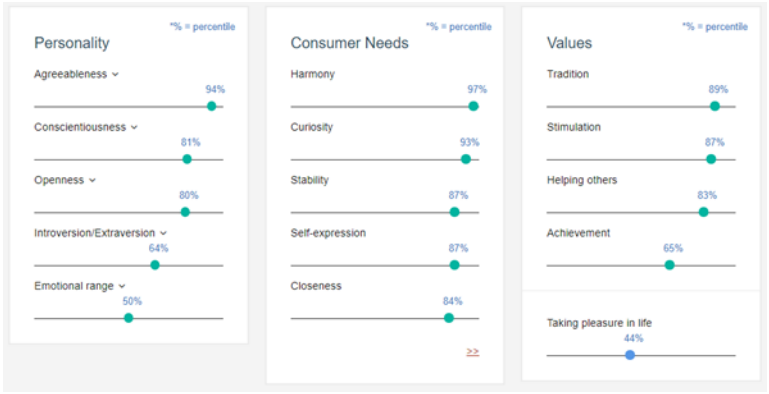
Commençons avec le gros morceau de ces APIs qui est lié au traitement du langage :

- **Assistant** (anciennement Dialog puis Conversation) permet de gérer des conversations scriptées pour des agents conversationnels, avec des arbres de décision. Ce genre d'outil est mis en œuvre depuis des années dans les systèmes de chat des sites de commerce en ligne. Les dialogues générés sont généralement limités car préprogrammés[682]. Les références d'IBM dans les chatbots réalisés avec Assistant sont nombreuses avec notamment celui du Crédit Mutuel, le Easy Button de Staples et Ask Mercedes.
- **Discovery** (anciennement Retrieve and Rank + Document Conversion) s'appuie sur le logiciel open source Apache Solr et permet de traiter les requêtes et questions en s'appuyant sur un mix de moteur de recherche et de machine learning. Il permet aussi de convertir tout document textuel (PDF, Word, HTML) pour les faire ingérer par les services de Watson. Cela sert à l'alimentation de bases de connaissances.
- **Natural Language Understanding** sert à l'extraction de données et de connaissances dans des textes.
- **Natural Language Classifier** permet de classifier automatiquement des données textuelles, issues en général de questions posées par des clients en langage naturel. Cela les associe aux questions d'une base existante de questions. Cela permet aussi de classifier des mails pour identifier des spams.

[681] Les requêtes REST permettent à une application html d'interroger un serveur web. Ce sont des requêtes http comprenant des GET et des POST et renvoyant le résultat.

[682] Voir ce tutoriel de développement de chatbot datant de début 2017, exploitant Conversation et d'autres briques logicielles d'IBM Watson : https://www.ibm.com/developerworks/library/cc-cognitive-chatbot-watson/index.html.

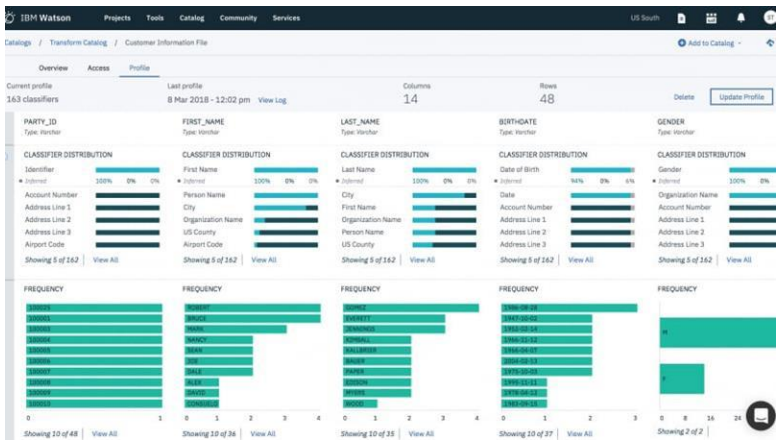
- **Personality Insights** analyse la personnalité d'un utilisateur à partir de ses textes. Ci-dessous l'analyse de la personnalité d'Oprah Winfrey basée sur ses tweets.



- **Language Translator** pour la traduction de textes. Il supporte 11 langues dont les grandes langues européennes et asiatiques.
- **Speech to Text** et **Text to Speech** pour compléter des chatbots en entrée et en sortie pour en faire des assistants personnels. L'ensemble supporte sept langues.
- **Tone Analyzer** analyse les émotions dans les textes comme dans les mails entrants de clients, dans les conversations des réseaux sociaux ou dans les chatbots. Cela renvoie une information sur le type d'émotion : colère, peur, …

Hors traitement du langage, il nous reste :

- **Visual Recognition** qui est destiné à toutes les applications de reconnaissance de l'image. Il est très utilisé dans le diagnostic dans l'imagerie médicale. Cette solution peut être exploitée dans des endroits inattendus comme avec la startup française 3D-minded, et son application mobile "Le Croqueur" qui identifie les chocolats de grands chocolatiers à partir d'une photo, sorte de Shazam du chocolat.
- **IBM Watson Studio** est l'environnement de travail qui permet à Watson de gérer ses projets d'IA. Il permet notamment d'accéder à des modèles de machine learning partagés dans la communauté des développeurs (vidéo) et sur quelques marchés (RH, marketing, supply chain, …). Studio gère tout le cycle d'un projet d'IA avec la récupération des données, la création des modèles de machine learning et de deep learning (avec un « network modeler » qui permet de créer des réseaux de neurones de manière graphique), la préparation des données, l'entraînement des modèles, leur déploiement dans le cloud puis leur gestion en production. L'environnement gère aussi la distribution des traitements, notamment sur des serveurs et des GPU de type Nvidia. L'outil fonctionne en mode texte, en mode commande ainsi qu'en mode interactif avec des menus tous prêts pour lancer les batches. Watson Studio est destiné aux experts métiers, aux data scientists et aux développeurs[683].
- Il comprend **Knowledge Catalog** qui est une sorte de gestionnaire des données qui vont ensuite alimenter les autres briques de Watson. C'est une sorte de « repository » des données récupérées qui peuvent être aussi bien quantitatives (tabulaires, bases de données) que textuelles. L'outil permet de faire des *business analytics* sur ces données.



[683] Voir Introducing IBM Watson Studio, mars 2018.

- **Machine Learning** recouvre les services en cloud de machine learning et deep learning qui sont exploitables dans Watson Studio. Les frameworks supportés comprennent TensorFlow, Spark ML, scikit-learn, XGBoost, Keras, Caffe (mais seulement la version 1), PyTorch et le modeleur d'IBM SPSS[684]. Watson Machine Learning gère le déploiement des solutions dans le cloud d'IBM.
- **IBM Watson Services for Core ML** permet de déployer des modèles entraînés de machine et deep learning pour les faire tourner sur iOS. Ce module était lancé en mars 2018.
- **Watson Explorer** permet de créer des solutions personnalisées d'analyse de données structurées et non structurées. Sachant que depuis début 2018, la nouvelle version de Cognos Analytics supporte maintenant des fonctions cognitives diverses comme le dialogue en langage naturel.
- **Content Hub** qui est en fait l'offre de CMS (content management system) d'IBM qui peut très bien tourner sans IA.

Watson est aussi décliné sur quelques marchés spécifiques au niveau applicatif avec **Watson Oncology** (cancérologie), **Watson Cybersecurity**, **Watson Virtual Agent** (des chatbots préconfigurés s'appuyant sur IBM Watson Assistant)

Créer une application Watson revient donc souvent à créer du code, du contenu et à réaliser un travail d'intégration pour créer un agent conversationnel intelligent ou des applications de machine learning ou de deep learning. Dans des approches verticales, il faut définir des scénarios de dialogues assez précis et avoir sous la main beaucoup de données exploitables, aussi bien structurées que non structurées.

D'où l'importance pour IBM d'avoir un écosystème de partenaires solutions à même de couvrir les besoins de divers marchés verticaux. Pour ce faire, IBM a lancé un programme partenaire assez classique qui comprend l'accès aux APIs, à une communauté, un programme d'accélération de trois mois et un catalogue de solutions pour promouvoir les partenaires. A ce jour, l'écosystème d'IBM Watson comprend environ plusieurs centaines de sociétés dont un bon nombre de startups. Le programme d'accélération porte surtout sur l'accompagnement technique mais donne aussi l'opportunité de pitcher son offre pour récupérer un part du fonds d'investissement de $100M créé pour l'occasion.

En plus de son écosystème, IBM développe l'activité de services pour prendre en main de bout en bout les projets de ses grands clients. Alors que l'équipe d'origine de Watson ne faisait que quelques personnes, elle comprendrait maintenant environ 10 000 personnes dans le monde, principalement des consultants, avant-vente et développeurs, dont 800 en France, y compris, un centre d'avant-vente et de support situé à Montpellier.

IBM a aussi ouvert un **centre de recherche** dédié à l'IOT et Watson à Munich associé à un investissement de $200M, probablement pluriannuel, et décliné Watson sur l'IOT avec des outils notamment dédiés à l'analytics et au machine learning.

Ces quelques milliers de personnes allouées à Watson sont un bon début mais encore peu au regard des plus de 200 000 collaborateurs d'IBM Services. La migration d'IBM vers un business "cognitif" suffisamment différentié des autres sociétés de services globales dans le monde est une course contre la montre. Et ces dernières ne se laisseront probablement pas faire, même si elles auront probablement quelque temps de retard à l'allumage et du mal à recruter (ou former, si on peut rêver) les talents en machine et deep learning.

Quid des tarifs de Watson ? Il serait fourni à coup de licence logicielle d'un prix supérieur au million de dollars, mais avec un tarif plus proche de ceux du cloud pour les partenaires.

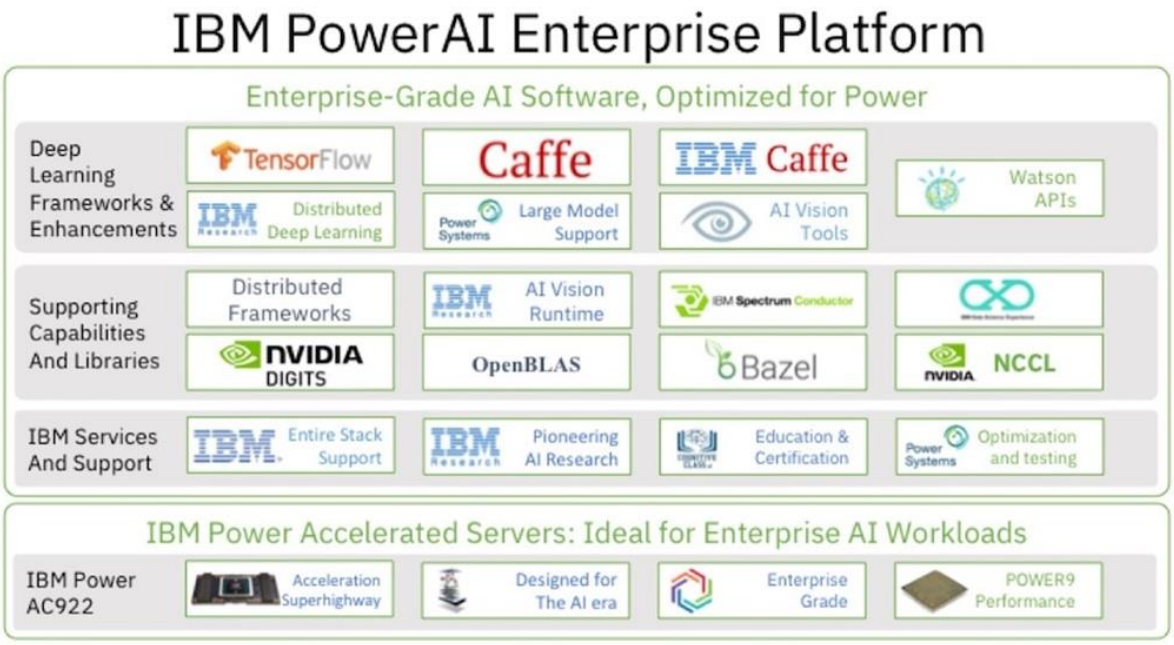
---

[684] Voir Deep Learning as a Service Now in IBM Watson Studio, mars 2018.

IBM prévoit de générer $10B de CA grâce à Watson d'ici une dizaine d'années. Ce qui ferait plus de 12% de son CA actuel.

Comment se positionne-t-il IBM par rapport à l'éventail des solutions du marché dans l'IA ? La comparaison est des plus difficiles. Il faut faire la part des choses entre briques prêts à l'emploi, outils de développement propres à IBM, frameworks tiers, outils de distribution des applications dans le cloud d'IBM. L'offre la plus voisine semble être celle de Microsoft. En Inde, WiPro propose aussi sa plateforme **Holmes** (*Heuristics and Ontology-based Learning Machines and Experiential Systems*) qui concurrence Watson[685].

IBM a aussi lancé **Power AI**, qui est le pendant infrastructure ouverte de Watson. En gros, c'est une offre matérielle et cloud générique capable de faire tourner des applications d'IA développées avec les outils du marché tels que Caffe, Theano et TensorFlow. Ici, Watson n'est plus spécifiquement de la partie.



Les données

IBM définit dans sa communication ce qu'est un bon projet pour Watson :

- Il doit traiter un **gros volume de données**. Makes sense !
- La solution doit permettre de **répondre rapidement** aux questions des utilisateurs, dans cette logique d'agent conversationnel fonctionnant en mode questions/réponses.
- La **variété des questions** traitées doit être grande grâce à une large palette de compréhension. Le système doit pouvoir traiter en profondeur les questions posées.
- Watson doit être en mesure **d'évaluer la validité des réponses**, avec un indice de confiance, comme il le faisait dans Jeopardy.

Les projets doivent être longs à conclure et à mener avec les grandes entreprises surtout si elles doivent mettre de l'ordre dans leurs données, comme ce fut le cas avec les projets de systèmes experts dans les années 1980.

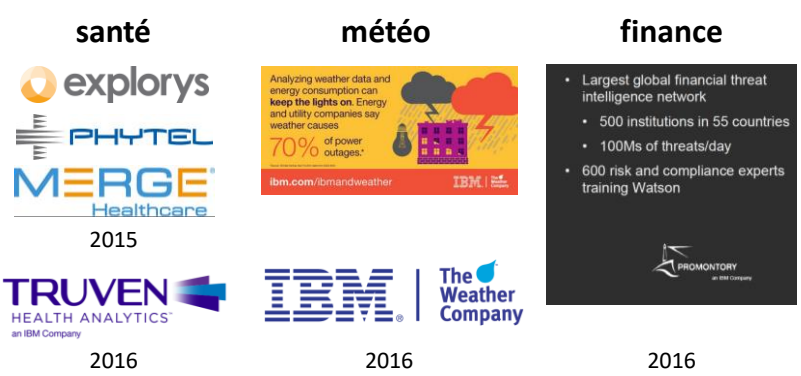
[685] Holmes est une plateforme d'IA hybride. C'est une boîte à outils voisine de celle d'IBM Watson avec de quoi créer un agent conversationnel, de la vision artificielle, du machine learning, à piloter drones et robots.

Ils ont probablement également des clients dans les secteurs militaires et du renseignement US qui ne donnent pas lieu à de la communication marketing. Finalement, les références sont maintenant bien plus nombreuses avec les partenaires éditeurs de logiciels qu'avec IBM en direct.

IBM est très dispersé verticalement mais avec un discours assez creux par secteur et relativement peu de références clients hors USA. Les outils marketing et la communication presse d'IBM répète le même discours générique sur Watson avec un zest de vertical. Quand aux déploiements annnoncés, il est toujours bien difficile d'évaluer s'ils sont sortis de la phase pilote.

Pour renforcer sa présence dans certains marchés verticaux, IBM a fait quelques acquisitions clés :

- Avec **The Weather Company** pour $2B en 2016, afin d'alimenter Watson avec des données météorologiques destinées à diverses applications comme pour identifier les risques météorologiques dans la définition de primes d'assurances dans l'immobilier. Et aussi pour prévoir le trafic de clients dans le retail.
- Avec **Promontory** en 2016 et ses outils de gestion de risques et de conformité permettant d'étoffer son offre dans la finance.
- Dans la santé, avec **Explorys** (2009, USA, $15M) avec sa plateforme de cloud dans la santé, **Phytel** (1996, USA, $22,5M) et sa solution de suivi de prise de traitements, **Merge Healhcare** (1987, USA) et ses outils de gestion d'imagerie médicale et **Truven Health Analytics** (2012, USA) et ses outils d'analytics.



  En 2018, IBM semblait dégresser les effectifs de certaines de ces acquisitions. Mais c'est un processus assez traditionnel après des acquisitions de startups et celles-là n'y ont pas échappé.

IBM a investi au moins $7B en acquisitions dans l'IA, bien plus que Google ne l'a fait. En plus des startups évoquées ci-dessous, il a notamment absorbé en 2014 la startup **Cognea** (2013, USA), créatrice d'un agent conversationnel, **AlchemyAPI** (2005, USA, $2M), une startup de deep learning d'analyse de textes et d'images, de reconnaissance de visages, de tagging automatique d'images acquise en 2015, et **IRIS Analytics** (2007, Allemagne), une startup d'analyse temps-réel dédiée à la détection de fraudes aux moyens de paiement, s'appuyant sur du machine learning.

Le matériel

Nous avons vu dans les parties concernant les processeurs neuromorphiques et quantiques qu'IBM était acteur intéressant avec d'un côté ses processeurs TrueNorth et de l'autre, ses premières expériences d'ordinateur quantique qui sont disponibles dans le cloud.

Tout ceci permet à IBM de conserver un peu d'avance dans sa capacité à produire des calculateurs de haute performance. Mais l'industrialisation à grande échelle est le bât qui blesse chez IBM. Pour que ces investissements soient rentables, il leur faudra générer du volume et trouver des débouchés pour ces composants. En effet, pour ce genre de technologie, rien ne dit que l'intégration verticale soit la meilleure approche. Surtout si la concurrence se structure de manière horizontale comme Intel le fait avec succès sur le marché des PC et des serveurs depuis 35 ans.

Fin 2017, IBM lançait ses serveurs Power9 censés être optimisés pour les applications d'IA. Leur processeur Power9 est un CPU assez classique mais très puissant. Il intègre le PCI-Express 4.0 et la technologie d'interconnexion CPU/GPU Nvidia NVLink.

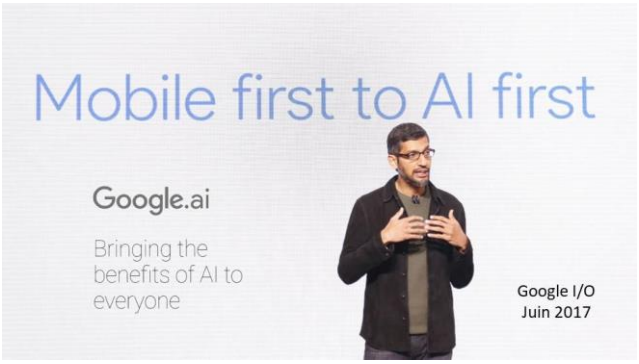
Tout ceci permet d'accélérer les transferts de données entre les composants, un point très important pour l'entraînement de modèles de machine learning. Ceci étant, cette architecture n'est pas idéale pour entraîner des modèles de deep learning.

Google

Depuis 2016, Google a mis la surmultipliée dans la communication autour de l'IA, battant IBM à ce jeu là. Presque tous les services de Google font appel à de l'IA : dans Google Search en général, dans la recherche d'images similaires de Google Search ou Google Photos, dans l'antispam de Gmail[686], dans l'assistant personnel Google Assistant, dans ses Google Car et même dans Android.

On en vient à oublier que Google était au départ un projet de création d'une vaste IA mondiale, démarré par un simple moteur de recherche. Ce serait devenu un véritable projet de société alimenté par un solutionnisme débridé[687].

S'y ajoute pour épater la galerie les performances multiples de leur filiale UK DeepMind, et notamment AlphaGo. Le CEO de Google démarrait la conférence Google I/O en juin 2017 en indiquant que la priorité numéro un de la société n'était plus la mobilité mais l'IA. En fait, dès l'an 2000, les fondateurs de Google considéraient que leur moteur de recherche n'était que la première brique d'une grande intelligence artificielle généralisée !



En matière d'IA, il est est relativement facile de différencier Google d'IBM. Les deux maitrisent des technologies logicielles sommes toutes assez voisines. La différence principale réside dans la manière de les mettre dans les mains des clients. Les deux ont des plateformes logicielles et cloud mises à disposition des développeurs et des startups.

Tandis qu'IBM fonctionne en mode projet et service avec les grandes entreprises, Google propose des services à plus de deux milliards d'internautes. Cela lui permet d'accumuler d'énormes volumes de données qu'il utilise pour entraîner « ses » IA, ce qu'IBM n'arrive à faire que sur certains marchés verticaux et via quelques acquisitions ciblées. Google a-t-il intérêt à copier IBM ? Pas vraiment. La profitabilité et la croissance de Google sont excellentes alors qu'IBM a une profitabilité de société de service et est en décroissance.

L'histoire

L'actualité abonde depuis 2014 d'acquisitions médiatisées de startups de l'IA par les grands acteurs du numérique, Google en premier.

Cela alimente quelques fantasmes sur leurs avancées qui sont quelque peu enjolivées. Google aurait, selon certains, acquis tout ce qui existerait de bien comme compétences dans l'IA. C'est évidemment une vue de l'esprit.

---

[686] L'IA d'antispam de gmail génèrerait seulement 0,05% d'erreurs. Elle exploite un système de deep learning réparti sur 16 000 CPU avec plus d'un milliard de connexions entre neurones.

[687] Voir la critique du modèle dans «La Vie après Google» ou l'espoir d'un monde nouveau, octobre 2018 au sujet de « Life after Google » de l'économiste George Gilder, qui entrevoit un monde décentralisé sauce Blockchain remplaçant le monde centralisé de Google.

Oui, Google a fait bien plus d'acquisitions dans le domaine de l'IA que les autres grands du numérique, mais rappelons-nous le côté très artisanal de ce secteur. Ce n'est pas parce que vous achetez quelques verreries de luxe que vous êtes le seul à savoir fabriquer des verres de luxe ! L'artisanat est très souvent un marché très fragmenté. On peut le constater au regard des effectifs des startups acquises. Ils sont en général très réduits, comme ils l'étaient d'ailleurs pour les acquisitions par Facebook de startups telles qu'Instagram, Whatsapp ou Oculus Rift, qui n'avaient par ailleurs aucun rapport avec l'IA.

L'acquisition la plus médiatisée de Google dans l'IA fut celle de **DeepMind** (2010, UK) en 2014 pour un montant record dans ce secteur de $625M. Et surtout, pour à peine une cinquantaine de personnes dont une douzaine de chercheurs en machine learning. Ce qui fait le chercheur à $50M, un record comparativement aux développeurs qui sont estimés à environ $1M à $2M pour des acquisitions de jeunes startups. DeepMind s'est depuis surtout fait remarquer en étant à l'origine des différentes moutures d'AlphaGo ou de réseaux de neurones comme PathNet.

Google avait auparavant mis la main sur la société de reconnaissance vocale **SayNow** (2005, USA, $7,5M) en 2011 puis sur **Viewdle** (2006, USA, $12M) et **PittPatt** (2004, USA) en 2012, qui faisaient tous les deux de la reconnaissance faciale et de mouvements. En 2013, ils mettaient la main sur le spécialiste des réseaux neuronaux **DNNresearch** (2012, Canada), et embauchaient ainsi le canadien Geoff Hinton, considéré comme le père du deep learning.

Ont suivi **Dark Blue Labs** (2014, UK) et **Vision Factory** (2014, UK), deux sociétés d'Oxford qui n'ont pas levé de fonds. S'y ajoutèrent le spécialiste de la traduction automatique **Quest Visual** (2009, USA), et celui de la reconnaissance de mouvements **Flutter** (2010, USA, $1,4M) qui a probablement enrichi l'offre logicielle de Dropcam, une startup de caméras de surveillance qui est dans le giron de Nest, une filiale d'Alphabet.

Depuis 2017, cette belle frénésie ininterrompue d'acquisitions dans l'IA semble s'être calmée.

Dans l'IA, on ne compte que celle d'**AIMatter** (2016, Biélorussie, $2M) avec sa plateforme de deep learning mobile de reconnaissance d'images et puis celle d'**Onward** (2015, USA, $120K), une plateforme de création de chatbot de service client.

Google a aussi acquis **Moodstocks** (2008, France, $500K) qui proposait une solution mobile de reconnaissance d'images, fournie sous la forme d'APIs et d'un SDK multi-plateforme. Cela semble être une acquihire.

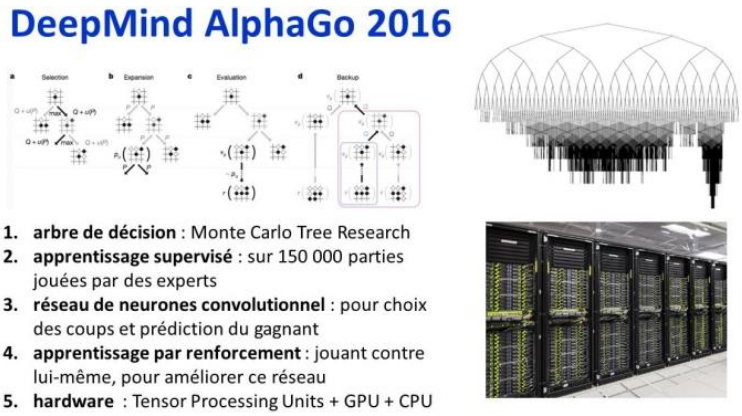
L'année 2014 avait vu Google/Alphabet acquérir une belle brochette de startups dans la robotique avec **Schaft** (robot humanoïde et bras articulé, japonais), **Industrial Perception** (robots industriels, spécialisé dans la vision 3D), **Redwood Robotics** (bras robotisés, issue du SRI et acquise un an après sa création), **Meka Robotics** (aussi dans les bras robotisés, qui avait contribué à la création de Redwood Robotics), **Holomni** (roues robotisées), **Bot & Dolly** (bras articulés à mouvements très souples servant aux tournages de cinéma), **Autofuss** (encore des bras articulés) et surtout **Boston Dynamics**, connu pour ses robots médiatisés doués de capacité de marche à quatre puis deux pattes mais que Google a cédé à **Softbank Robotics** en juillet 2017.

Schaft a été également cédé à Softbank Robotics la même année ! Bref, la stratégie robotique « attrape tout » de Google est à prendre avec des pincettes.

Que deviennent toutes ces acquisitions ? Tout ce qui relève du traitement des images et du langage s'est retrouvé dans les services de Google, notamment mobiles. La robotique ? Elle a débouché sur aucune application commerciale pour l'instant car ces technologies sont toujours en phase de gestation ou destinées à des marchés de niche, ou aux usages internes de Google comme pour les véhicules utilisés pour cartographier les rues. Et Google ne cherche pas à concurrencer les leaders de robots industriels (ABB, Fanuc, etc)[688].

Google fait des progrès réguliers dans le traitement des images, comme avec **PlaNet** qui identifie à quel endroit ont été prises des photos d'extérieur ou pour compter les calories dans des **photos de plats cuisinés**. Google utilise aussi beaucoup d'IA sensorielle pour faire évoluer les fonctions de conduite automatique de ses Google Car.

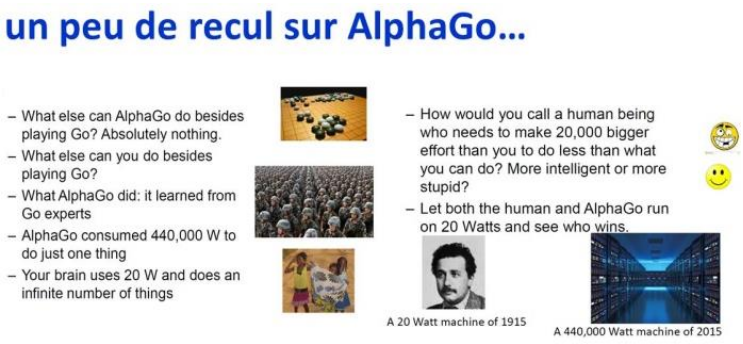
L'IA googlelienne a connu un sursaut de médiatisation début 2016 avec la victoire de la solution AlphaGo contre le champion du monde coréen de Go, Lee Sedol construite par une équipe d'une vingtaine de personnes de sa filiale DeepMind (**vidéo de la première partie**). Ces victoires ont été présentées comme des étapes importantes des progrès de l'IA, faisant écho à la victoire de Deep Blue aux échecs contre Gary Kasparov en 1997.



La différence ? Le jeu de Go est plus difficile à simuler car la combinatoire de jeu est bien plus grande qu'aux échecs. AlphaGo ne peut donc pas compter que sur la force brute.

Il doit combiner plusieurs méthodes pour être efficace : éliminer des options de jeu inutiles via le "Monte Carlo Tree Search" ou MCTS et exploiter une base de jeux permettant d'identifier des tactiques gagnantes. Il réalise ensuite un apprentissage supervisé à base de deep learning en exploitant 150 000 parties connues. Il choisi ses coups avec un réseau de neurones convolutionnel. Il fait de l'apprentissage par renforcement en jouant contre lui-même. Et l'ensemble exploite la puissance machine de GPU et de TPU que nous avons vu dans la partie consacrée aux processeurs neuromorphiques. En 2017, AlphaGo Zero faisait progresser la discipline en s'entrainant par lui-même et en étant plus économe en ressources. Nous l'avons déjà évoqué.

Dans "Artificial Intelligence and the Singularity" publié en 2016, Piero Scaruffi se faisait un malin plaisir de relativiser cette victoire (*ci-contre*) en rappelant la consommation d'énergie du système par rapport aux 20W du cerveau humain ! On pourrait ajouter que depuis plus de 30 ans, n'importe quel tableur gagnerait haut la main toute partie contre les champions du monde du calcul mental !



[688] Voir Google robotics: A review - The Robot Report, octobre 2017, et Google Has Made a Mess of Robotics de Mark Bergen et Joshua Brustein, octobre 2017, qui racontent les déboires de Google dans la robotique.

La performance a été documentée dans un article publié dans la revue Nature en janvier 2016[689]. Un peu vexés, les coréens ont d'emblée lancé un plan de financement public de 765M€ dans l'IA sur cinq ans avec les géants comme Samsung, LG, Hyundai et SK Telecom[690]. C'était suivi par un nouveau plan de $2B lancé en mai 2018[691].

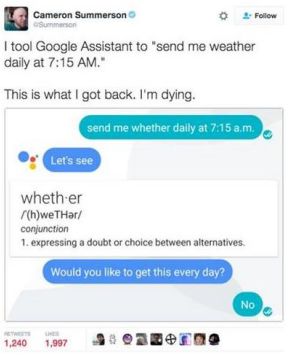
En 2015, Matthew Lai développait **DeepChess**, un système de deep learning avec renforcement qui gagnait aux échecs en apprenant lui-même à optimiser son jeu en moins de 72 heures sur un simple PC. Il était recruté par DeepMind début 2016 et il a contribué aux évolutions d'AlphaGo à partir de ce moment-là !

Tout cela faisait en tout cas une excellente publicité pour DeepMind dont les solutions de machine learning ont heureusement d'autres applications comme la **curation de médias**, même si elles font moins parler d'elles car elles ressemblent de près à ce qu'IBM fait déjà dans la santé avec Watson.

Leur **DQN** est un réseau neuronal profond doté de capacités d'auto-apprentissage et **DeepMind Health** donne lieu à une collaboration avec la NHS britannique dans l'application Streams de détection de blessures aux reins dans les urgences. En novembre 2018, Alphabet annonçait regrouper ses activités disparates dans la santé sous l'ombrelle de Google Health.

Les logiciels

Google utilise l'IA pour enrichir ses propres offres grand public, que ce soit autour de son moteur de recherche multifonctions ou de business plus périphériques d'Alphabet (santé, IoT, automobile). On la rerouve aussi dans Google Assistant et Google Home, ces agents conversationnels pilotables à la voix et au clavier.



C'est aussi un grand fournisseur de platteformes de développement en open source ou pas, et dans l'embarqué ou en cloud.

Il publie régulièrement de nombreuses **APIs de services d'IA dans le cloud** pour les développeurs (Google Cloud Machine Learning Engine). Google est aussi à l'origine de la bibliothèque de machine et deep learning **TensorFlow** qui est très couramment utilisée par les startups de l'IA. Comme IBM, nombre de ses services couvrent le traitement du langage, y compris la traduction, ainsi que la vision artificielle. Cette panoplie couvre une majeure partie des besoins de créateurs d'applications à base d'IA.

---

[689] Voir Mastering the game of Go with deep neural networks and tree search, janvier 2016 (37 pages).

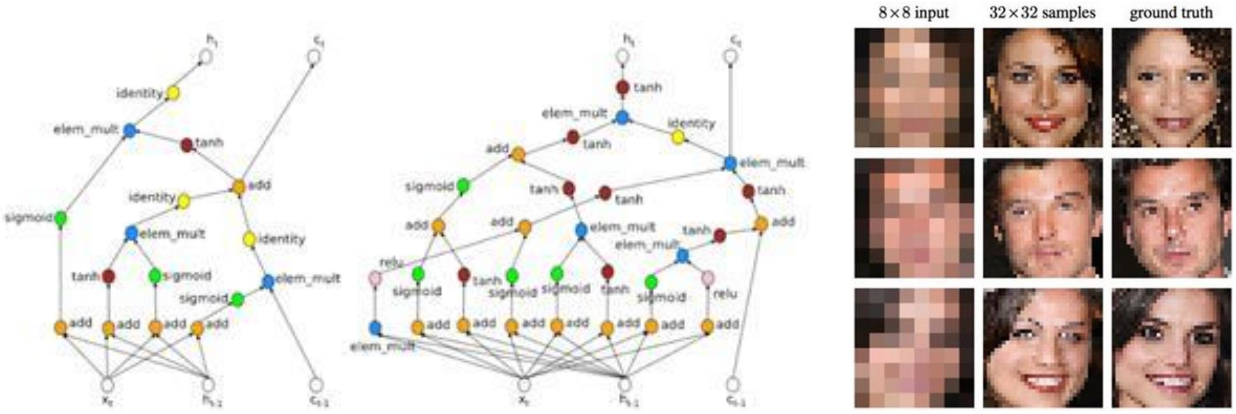
[690] Voir Intelligence artificielle : la Corée (vexée d'avoir perdu au jeu de go) annonce un plan de 765 M€, de Julien Bergounhoux, mars 2016. En mai 2017, une version améliorée d'Alpha Go battait le champion chinois Ke Jie.

[691] Voir South Korea Aims High on AI, Pumps $2 Billion Into R&D, mai 2018.

Mais Google regorge de développeurs et d'équipes projets en tout genre. Nous avons par exemple les projets **Google Brain**[692] lancés en 2011 par une petite équipe de chercheurs qui comprend Jeff Dean, Greg Corrado, Andrew Ng et Geoff Hinton depuis 2013.

Cette équipe est à l'origine de systèmes de chiffrement évolutifs publiés en octobre 2016[693] et d'un étonnant programme d'amélioration d'images pixellisées publié en en février 2017, exploitant des images de 8x8 pixels[694] pour augmenter leur résolution à 32x32 pixels. L'image du milieu est celle que l'IA de Google reconstitue à partir de celle de gauche. C'est impressionnant. Mais attention au fait que les images de départ semblent être des versions basse résolution de la base d'entraînement.

En mai 2017, ils publiaient aussi d'étonnants travaux montrant comment ils pouvaient utiliser le machine learning pour améliorer l'architecture d'un réseau de neurones qui est d'habitude créée manuellement (*illustrée ci-dessous à gauche*[695]).



Les développeurs de Google sont à l'origine d'avancées régulières dans les réseaux de neurones de reconnaissance d'image. C'est le cas de **Facenet** qui améliore les techniques de reconnaissance de visages, entraîné sur 260 millions d'images et efficace à 86% en 2016. La méthode ? Une variante de réseau de neurone convolutionnel[696] (*ci-dessus à droite*). L'équipe de Google Brain est aussi à l'origine d'améliorations diverses de Google Translate[697].

Parfois, la communication de Google DeepMind en fait un peu trop, comme avec Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents, janvier 2018 (28 pages). Il s'agit d'un toolkit qui évalue les capacités cognitives d'une IA à base d'agents (UNREAL = UNsupervised REinforcement and Auxiliary Learning) en s'appuyant sur les mêmes modèles psychophysiques que pour l'évaluation des capacités cognitives humaines à se déplacer dans l'espace pour atteindre un objectif. Il évalue ici les IA d'apprentissage par renforcement dans le contexte de la détection visuelle de formes. L'évaluation montre que l'apprentissage par renforcement a une capacité à planifier ses actions sur le long terme (vidéo). Tout cela est sympathique mais c'est de l'IA très très étroite. L'agent UNREAL s'appuie sur un réseau convolutionel (CNN) couplé à un réseau à mémoire (LSTM).

---

[692] Google Brain est distinct de Deep Mind qui est basé au Royaume Uni, reste relativement indépendant après son acquisition en 2014.

[693] Qui là aussi a beaucoup fait fantasmer avec « l'IA qui créé son propre langage que les hommes ne comprennent pas ». Voir Google taught artificial intelligence to encrypt messages on its own de Dave Gershgorn, octobre 2016.

[694] Voir Pixel Recursive Super Resolution, février 2017. Qui rappelle le scénario du film « No way out » avec Kevin Costner, sorti en 1987. Il faut préciser que le système est entraîné avec les images de la dernière colonne.

[695] Voir Using Machine Learning to Explore Neural Network Architecture, mai 2017.

[696] Voir FaceNet: A Unified Embedding for Face Recognition and Clustering, juin 2015. L'un des trois auteurs, James Philbin, a quitté Google en 2015. Il est depuis le directeur de la vision artificielle de **Zoox** (2014, $290M), une startup ultra-bien financée qui veut devenir un opérateur de service de véhicules autonomes.

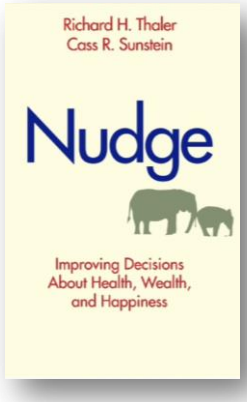
[697] Voir Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016.

Lors de la conférence Google I/O de mai 2018, Google présentait quelques autres nouveautés en matière d'IA, surtout du traitement du langage : des évolutions de Google Image dont une fonction de colorisation, reprenant des algorithmes génératifs déjà bien connus, l'auto-remplissage d'emails avec la fonction Smart Compose (vidéo) et Google Duplex (vidéo) pour faire prendre des rendez-vous au téléphone par un assistant vocal.

Autant la première démonstration est plausible et s'appuie sur des modèles probabilistes de réseaux de neurones à mémoire (LSTM ou variante), autant la seconde est sujette à caution tant que l'on n'a pas pu la tester dans une grande variété de cas.

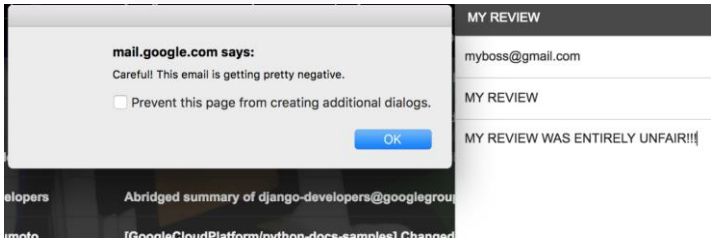
On peut aussi citer le plugin Chrome DeepBreadth qui prévient l'utilisateur en train de rédiger un email un peu incendiaire pour lui recommander de réfléchir avant de l'envoyer. Ce type de fonctionnalité était évoquée de manière prémonitoire dans l'ouvrage Nudge de Richar Thalet et Cass Sunstein en 2008 (voir *ci-dessous*).

Enfin, la version à venir d'Android, Android P contient des « actions » et « slices », des fonctions pour les développeurs qui permettent de prévoir la prochaine action de l'utilisateur (vidéo). Mais ces prévisions sont à faire par les développeurs. Android P ne fournit que les briques logicielles et interfaces de programmation permettant de les intégrer dans Android. C'est à se demander ce qu'il reste du libre arbitre comme l'évoque à juste titre Gaspard Koenig !

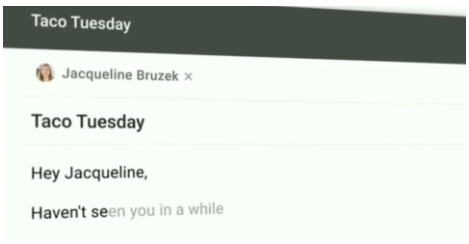


We propose a Civility Check that can accurately tell whether the email you're about to send is angry and caution you, "WARNING: THIS APPEARS TO BE AN UNCIVIL EMAIL. DO YOU REALLY AND TRULY WANT TO SEND IT?" (Software already exists to detect foul language. What we are proposing is more subtle, because it is easy to send a really awful email message that does not contain any four-letter words.) A stronger version, which people could choose or which might be the default, would say, "WARNING: THIS APPEARS TO BE AN UNCIVIL EMAIL. THIS WILL NOT BE SENT UNLESS YOU ASK TO RESEND IN TWENTY-FOUR HOURS." With the stronger version, you might be able to bypass the delay with some work (by inputting, say, your Social Security number and your grandfather's birth date, or maybe by solving some irritating math problem!).*

Nudge 2008



Google DeepBreath Chrome plugin 2017



Gmail Smart Compose 2018

Les données

Il va sans dire que la puissance de Google vient de la quantité astronomique de données qu'ils accumulent sur les faits et gestes Internetiens et même dans le monde physique de millions d'utilisateurs. En gros, Google sait ce que l'on recherche (Search), où l'on est et où l'on va (Android, Maps), quels moyens de transport on utilise (Maps, Android), ce que l'on échange avec les autres (Gmail), le temps que l'on passe sur tel et tel écran, et plus rarement, ce que l'on cherche et regarde à la TV (Google TV).

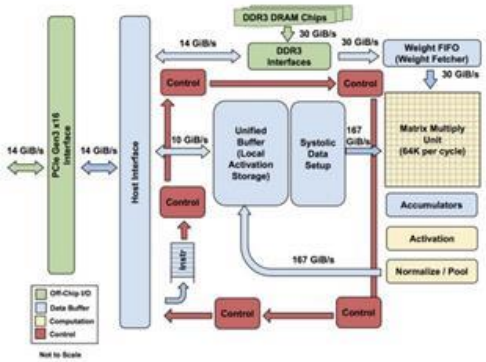
Ils ont la compétence pour stocker, indexer et gérer ces données dans le cloud. A partir de là, ils peuvent tester tout un tas d'idées sans grandes limites ! Mais il ne faut pas non plus exagérer leur puissance. Ils n'accèdent pas aux données métier des entreprises et en particulier à ceux des infrastructures. Ils n'ont pas accès aux données bancaires, d'assurances, de la distribution ou dans la san-

té, même si dans ce dernier cas, ils fournissent les ressources techniques de calcul en cloud à des chercheurs en génomique à divers laboratoires de recherche.

Le matériel

Google a développé ses propres processeurs neuromorphiques en 2015/2016, le TPU[698]. Ce sont des processeurs programmables capables de gérer des réseaux de neurones « fully connected » (couches de neurones interconnectées) ainsi que les premiers étages de réseaux de neurones convolutionnels (nécessitant des multiplieurs de matrices). Ces TPU ont été utilisés par AlphaGo de DeepMind et gèrent bien d'autres briques d'AI de Google. Il est cependant difficile de savoir ce qui est réellement en production dans le cloud de Google avec ces TPUs.

En 2018, ils annonçaient leur troisième génération de TPU. Ils sont maintenant positionnés comme de véritables serveurs multi-processeurs. Nous les avons déjà décrits dans la rubrique sur les processeurs neuromorphiques.



Google n'a pas l'air de prévoir la commercialisation de ces processeurs et serveurs. C'est une technologie fabriquée en volume en ASIC destinée à équiper leurs propres datacenters. Mais comme Google est le premier consommateur de serveurs au monde, ils ont les économies d'échelle qui le permetttent.

Pour ce qui est des ordinateurs quantiques, Google a commencé par tester avec la NASA les ordinateurs à recuit quantique du Canadien **D-Wave**. Google met aussi au point ses propres ordinateurs quantiques à qubits supraconducteurs – effet Josephson. Ils annonçaient en mars 2018 avoir atteint 72 qubits, mais sans les avoir démontrés dans une machine opérationnelle. Ils ont sinon déjà créé deux frameworks de développement : **Cirq** pour le développement à bas niveau et **OpenFermion** pour le développement d'applications dans le domaine de la simulation chimique.

Google est sinon concepteur d'objets connectés intégrant des logiciels d'IA, qu'il s'agisse de **Google Home**, des smartphones **Pixel** ou des futures **Google Car**, qui pourraient à terme être fabriquées par des constructeurs automobiles, Google fournissant l'électronique, les logiciels (sous Android) et les solutions en cloud associées. Cette activité est maintenant intégrée dans la filiale **Waymo** du groupe Alphabet dont fait partie Google[699].

Amazon

Amazon est un acteur clé de l'IA dans les entreprises via son offre intégrée de cloud sous la bannière des **Amazon Web Services** (AWS). Amazon est le leader mondial des services génériques en cloud, utilisé largement par les entreprises tout comme par les startups. Il devance de loin Microsoft Azure qui est le second sur ce marché.

---

[698] Voir First in-depth look at Google's TPU archtecture, de Nicole Hemsoth, avril 2017.

[699] A noter que le Directeur de l'Engineering de Waymo, la filiale pour véhicules autonomes de Google, est le Français Sacha Arnoud, issu de l'X.

amazon

offre	forces	faiblesses
Lex, Rekognition, Polly Alexa (pf NLP) Echo (device) Spark (big data distribué) AWS (cloud services) DSSTNE (deep learning) Kiva (robots)	services grand public données utilisateurs utilisateur intensif de ML dans ses propres services leader mondial du cloud logistique médias et contenus	faible différenciation logicielle dans l'IA pas habitué à évangéliser les développeurs comme IBM, Google et Microsoft

Amazon est comme tous les GAFAMI un acquéreur de startups régulier, mais pas très actif du côté de l'IA. On peut noter celle du spécialiste des robots d'entrepôts **Kiva** (2003, USA, $18M) en 2012 pour $775M, d'**Ivona** (2004, Pologne), spécialiste du text to speech, acquis en 2013, d'**Angel.ai** (2015, USA, $8M) en septembre 2016, créateur d'un chatbot généraliste qui a certainement du les aider à améliorer Alex et de **harvest.ai** (2014, USA, $2,74M) détection de failles de sécurité, acquis en 2017.

Kiva, robots d'entrepôts, acquis en 2012 pour $775M

Safaba, traduction automatique, acquis en 2015

Orbeus, recherche d'images, acquise en 2016

Angel.ai, chatbot généraliste acquis en 2016

harvest.ai, détection de failles de sécurité, acquis en 2017

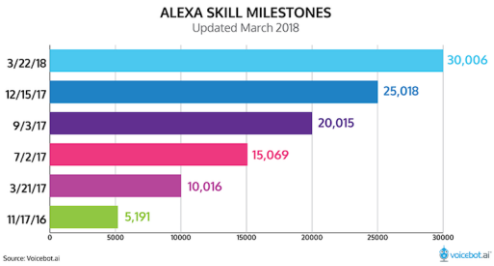
Ivona, acquis en 2013

Les logiciels

Vu du grand public, Amazon est aussi présent dans l'IA via les nombreuses déclinaisons d'**Amazon Echo** et le service en cloud de dialogue en langage naturel **Alexa** qui l'équipe et qui est très largement utilisé par l'écosystème des objets connectés.

Il était quasiment devenu le standard par défaut des objets commandables à la voix introduits au CES 2017. On le trouvait ainsi supporté par un nombre incalculable d'objets connectés : radioréveils, robots chez **UBTech**, lampes connectées, copycats de l'Amazon Echo chez **Lenovo**, routeur chez **Linksys**, dans l'électroménager chez **LG Electronics** ou chez **GE**, dans les aspirateurs robots de **Samsung** et **Neato** et même chez **Ford** qui le supportera dans ses véhicules pour commander leur démarrage à distance et gérer les parcours (vidéo).

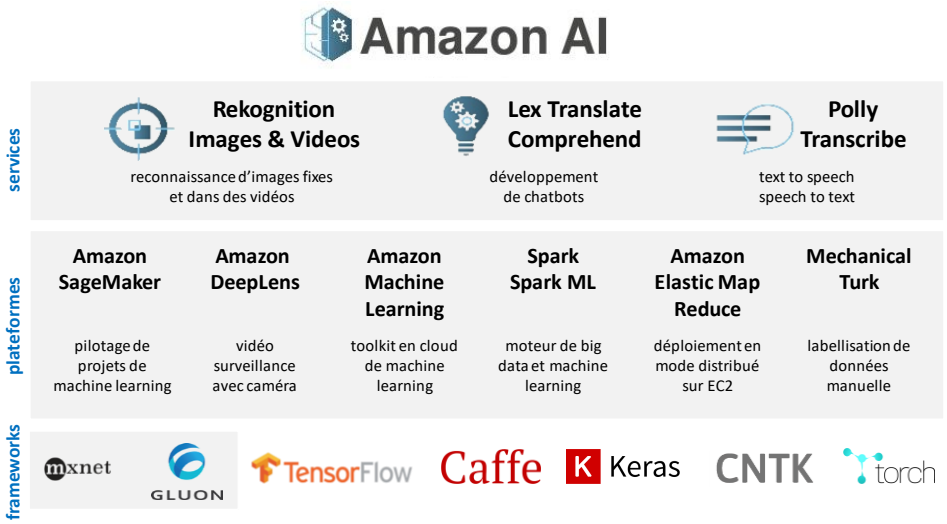
Il y aurait plus de 40 000 applications pour Alexa, dénommées *skills*, rien que pour le marché américain [700]. Depuis début 2018, Google a rattrapé du terrain face à Amazon avec son Assistant et ses hauts parleurs connectés. Amazon domine encore ce marché mais sa part s'érode inexorablement.



[700] Source : https://www.voicebot.ai/amazon-echo-alexa-stats/.

L'offre d'APIs d'AI pour les développeurs d'applications en cloud comprend, outre Alexa :

- **Amazon Rekognition**, une fonctionnalité d'analyse d'images fixes et de vidéos à base de deep learning qui permet d'identifier des objets, de les tagger, d'éliminer des contenus illicites, d'analyser les expressions dans les visages et de les reconnaître.
- **Amazon Polly** est une solution de text-to-speech réaliste lancée fin 2016 (vidéo et conférence technique) avec un choix de 47 voix dans 25 langues**. Amazon Transcribe** réalise la fonction inverse, le speech to text.
- **Amazon Lex** et **Comprehend** sont les moteurs de gestion de conversations d'Alexa.
- **Amazon Translate**, est un moteur de traduction issu de la startup américaine **Safaba Translation Solutions** (2009, USA) acquise en 2015.



Du côté des couches basses, nous avons :

- **Amazon SageMaker** est un environnement complet de gestion du cycle de gestion des projets de machine learning. Il lance toutes les briques intermédiaires (vidéo). On charge les données, entraîne des modèles, les teste puis les met en production. Les services de SageMaker se pilotent à partir d'un environnement de développement (notebook) en mode web dénommé **Jupyter**.
- **DeepLens** est une solution complète de reconnaissance d'images intégrant une caméra et les logiciels de traitement des vidéos qu'elle génère. La caméra utilise un processeur Atom tournant sous Linux avec une puissance de 100 Gflops, 8 Go de mémoire et 32 Go de stockage. Elle capte les photos avec 4 mpixels en MJPEG et les vidéos Full HD en H.264. La caméra a probablement été conçue par les équipes de la startup **Blink** (2009, USA, $5,8M) acquise en décembre 2017. La partie serveur s'appuie sur Amazon Rekognition Video.
- **Amazon Machine Learning** consolide les outils de création et exécution de modèles de machine learning.
- **DSSTNE** (Deep Scalable Sparse Tensor Network Engine, ou « destiny ») qui permet de créer des modèles de machine learning et de deep learning faciles à déployer sur GPU… Nvidia en général.
- **Spark** pour la distribution de traitements sur serveurs, logiciel de la fondation Apache et **SparkML** ou Spark MLlib, une bibliothèque qui permet de distribuer des traitements de machine learning ainsi que **BigDL**, une bibliothèque de deep learning. Tous ces logiciels sont open source ! Les entreprises payent les ressources en cloud pour les héberger.
- **Amazon Elastic MapReduce**, qui permet de déployer des applications Spark et Hadoop sur plusieurs serveurs, notamment pour le traitement de gros volumes de données.

Au niveau des frameworks :

- **Mxnet** est un framework complet de la même catégorie que PyTorch et TensorFlow, et optimisé pour fonctionner sur le cloud d'Amazon.
- **Gluon** est un framework de plus haut niveau au-dessus de Mxnet. C'est en quelque sorte l'équivalent de Keras pour Tensorflow.

Les données

Accessoirement, Amazon est le leader mondial du commerce en ligne et la part qu'il représente dans ce marché est en croissance, surtout aux USA où il captait 46% du marché en 2016.

C'est lui qui possède le plus gros inventaire de produits dans son catalogue, qui est estimé à plusieurs centaines de millions de produits, notamment via les offres intégrées dans sa place de marché. En conséquence de quoi, comme Google, il dispose d'un beau pactole de données pour analyser les comportements des Internautes dans leur casquette de consommateurs. Il dispose aussi de données sur la consommation culturelle via ses services Prime Video, ses tablettes Kindle et sa box TV Fire, surtout aux USA.

Amazon est donc un gros utilisateur et de longue date de techniques de machine learning pour optimiser tout son processus de vente et de logistique. Il les utilise pour planifier la demande et gérer au plus près les stocks, pour définir les prix, les offres de livraison, pour la recommandation de produits, la détection de fraudes et de contrefaçons, pas toujours parfaite d'ailleurs. Cela explique probablement pourquoi Amazon est l'un des plus gros recruteurs aux USA de spécialistes de machine learning[701].

Le matériel

Contrairement à Google, Amazon n'a pas encore développé ses propres chipsets d'IA pour serveurs. Depuis fin 2017, ses serveurs EC2 permettent l'accès à des GPU **Nvidia** V100, jusqu'à par paquets de 8 équivalents aux serveurs Nvidia DGX1.

Par contre, Amazon serait en train de développer ses propres chipsets embarqués pour faire fonctionner Alexa en s'appuyant sur les compétences de la startup **Anapurna** (2011, USA) acquise en 2015.

Microsoft

Microsoft est un acteur de poids dans les infrastructures informatiques des entreprises et son arsenal logiciel dans l'IA est équivalent sur de nombreux points avec celui de ses concurrents tels qu'IBM et Google.

De manière assez classique, son offre couvre le traitement du langage, de la vision artificielle, des données et la gestion des connaissances, ainsi que tout ce qui permet de gérer des modèles de machine et deep learning. L'éditeur est plus discret dans sa communication autour de l'IA et son marketing produit est moins efficace, tout du moins comparativement à IBM et Google. Par contre, l'éditeur a mis toute la voilure sous le vent de l'IA depuis 2017.

[701] Voir IBM (IBM) Creating Shareholder Value with AI? Not so Elementary, My Dear Watson de Jefferies, juillet 2017.

Microsoft met aussi le paquet pour promouvoir une “IA responsible et éthique”, comme dans l’ouvrage The Future Computed, Artificial Intelligence and its role in Society, 2018 (152 pages) créé par l’équipe des juristes et de relations institutionnelles de Microsoft. En complément, le Guide de survie de l’IA publié par Microsoft France en 2018 est un catalogue d’études de cas par marchés verticaux réalisées avec des partenaires services et logiciels.

En France, ils ont lancé une école de l’IA de formation sur 7 mois de jeunes et personnes au chômage, opérée par Simplon.co, avec comme parrains des deux premières promotions Aurélie Jean et Benoit Raphaël. A Station F, leur IA Factory accélère des startups de l’IA. Les personnes formées passent ensuite 12 mois en alternance chez des partenaires de Microsoft. Cela concerne deux douzaines de personnes par fournée.

En mars 2018, Microsoft annonçait devenir partenaire de l’**Institut PRAIRIE** (PaRis Artificial Intelligence Research InstitutE) créé conjointement par le CNRS, l’Inria, et PSL et des entreprises privées, dont Amazon, Google, Facebook, Criteo, Faurecia, Naver Labs, Nokia Bell Labs, PSA, Suez et Valeo. Cela sera donc un temple de la recherche partenariale en IA. Cela complète le laboratoire de recherche conjoint entre Microsoft et l’Inria situé à Orsay et lancé en 2005.

Les logiciels

Microsoft a ceci de commun avec IBM qu’il entretient depuis des décennies de grandes équipes de recherche fondamentale et particulièrement investies dans les différents champs de l’IA.

Créé en 1991, **Microsoft Research** occupe plus de 1000 chercheurs répartis dans le monde, et y compris en France, dans un laboratoire commun monté à Orsay avec l’INRIA. La principale équipe européenne est située à Cambridge au Royaume-Uni. L’équipe de Microsoft Research a été associée aux équipes technologiques en charge de l’IA en 2017. L’ensemble comprend 8000 collaborateurs.

Microsoft Research emploie un nombre record de prix Nobel et de scientifiques ayant gagné la médaille Fields. Cela n’en fait pas pour autant les initiateurs de business significatifs pour Microsoft. Tout au plus sont-ils à l’origine de nombreuses innovations incrémentales qui ont alimenté les produits phares de l’éditeur. Le correcteur orthographique qui souligne les mots dans Word était ainsi sorti de ces laboratoires en 1995. Cela permet de relativiser le rôle de la recherche pour dominer une industrie. Apple qui n’a pas formellement de laboratoire de recherche domine ainsi le secteur du mobile, en compagnie de Google ! Chez Google, la frontière entre recherche et développement est plus floue. Microsoft et Google sont en fait très proches dans leur équilibre de R&D.

Les activités de Microsoft Research dans le machine learning sont imposantes avec **plusieurs dizaines d’équipes projets** impliquées. Dans les projets, on trouve les grands classiques qui portent sur l’amélioration de la reconnaissance de la parole et des images et notamment le tagging automatique de vidéos. Et puis, en vrac, un agent conversationnel détectant des troubles psychiatriques (**DiPsy**), un outil de reconnaissance de chiens originaire de Chine qui fonctionne à l’échelle individuelle, pas à celui de la race (**Dog Recognition**) et un outil de tri de pièces de monnaie pour les réfractaires aux Blockchains (**Numiscan**).

Les équipes de Microsoft Research sont à l’origine d’avancées comme le système de dialogue en langage naturel **Cortana**. Comme nombre de technologies d’IA proviennent de MSR chez Microsoft, l’éditeur se retrouve à mettre systématiquement en avant les travaux de ses chercheurs, parfois un peu trop au détriment des équipes produit classiques.

Microsoft qui est maintenant résolument tourné vers le cloud fait tout de même quelques acquisitions de startups pour accélérer son “time to market” dans l’IA ou dans la périphérie de l’IA. Les équipes de recherche fondamentale travaillent en effet sur des domaines où le risque est plus scientifique et technique que marché tandis que les startups sont censées œuvrer une un risque marché.

Le risque est même parfois émotionnel et dans l’image, comme l’a montré le robot conversationnel **Tay** qui s’est mis à tenir des propos nazis et a été débranché. Tay était sorti de Microsoft Research et ses propos relevaient d’un apprentissage supervisé non filtré !

Tay a été remplacé en avril 2017 par un autre chatbot au doux nom de **Zo** qui est intégré dans la messagerie instantanée **Kik**. Zo est une version anglaise d'un chabot chinois de Microsoft dénommé **Xiaoice**. Mais Kik n'est pas très trendy chez les Internautes !

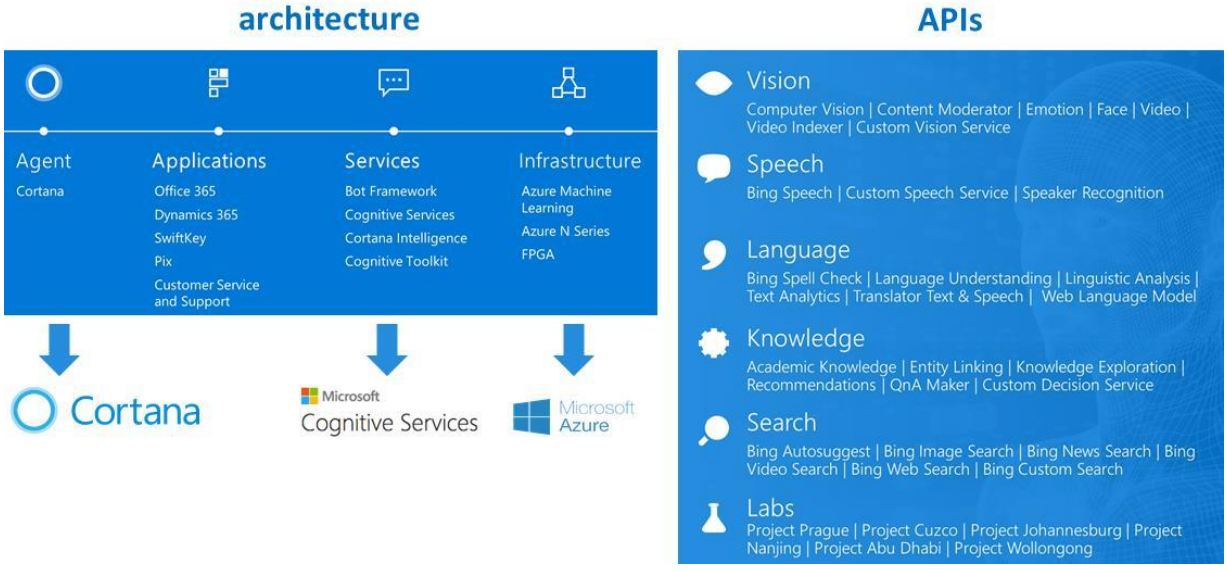
Les acquisitions dans les startups de l'IA sont relativement peu nombreuses chez Microsoft. En 2015, l'éditeur avait mis la main sur **Prismatic** (2010, USA, $15M), un agrégateur de news s'appuyant sur du machine learning, ainsi que **Double Labs** (2013, USA), une application Android de notification elle aussi basée sur du machine learning.

En 2016, c'était au tour de **Revolution Analytics** (2007, USA ? $38,7M), qui faisait de l'analyse prédictive s'appuyant sur le langage open source R, acquise en 2016. Un moyen de s'attirer un écosystème de développeurs ! Toujours en 2016, **Swiftkey** (200, UK, $21,6M), un logiciel de clavier virtuel mobile qui s'appuierait lui aussi sur du machine learning.

En 2017, Microsoft faisait l'acquisition de **Genee** (2014, $1,45M) un gestionnaire d'agenda virtuel et du Canadien **Maluuba** (2011, $8,2M), un spécialiste du deep learning appliqué au traitement du langage qui travaille sur l'AGI (Artificial General Intelligence), dont les équipes se sont faites remarquer en faisant gagner leur solution au Pac-Man avec leur technique Hybrid Reward Architecture[702] (vidéo).

En juin 2018, **Bonsai** (2014, USA, $13,6M) apportait dans la corbeille son BRAIN (Basic Recurrent Artificial Intelligence Network), un système de machine learning avec un haut niveau d'asbtraction, destiné surtout aux systèmes embarqués, et doté de capacités d'apprentissages par renforcement, comme pour l'entraînement de robots à saisir des pièces. La startup avait été créé par deux anciens ingénieurs de l'éditeur, Mark Hammond et Keen Browne. Retour au bercail ! En septembre 2018, c'était enfin le tour de **Lobe** (USA) de passer sous le giron de Microsoft. La startup a développé un système de création visuelle de modèles de deep learning à base de glisser-déplacer.

Il n'empêche que l'éditeur a bien compris les enjeux de l'IA et cherche à se positionner comme fournisseur de plateforme d'IA pour les développeurs, avec le "Conversation As a Platform" et le "Microsoft Bot Framework", qui rappellent dans leur structure l'offre des APIs d'**IBM Watson**. Il a été annoncé lors de la conférence Build qui s'est tenue à San Francisco en avril 2016 (voir les vidéos de keynotes du **premier jour** et du **second jour**).



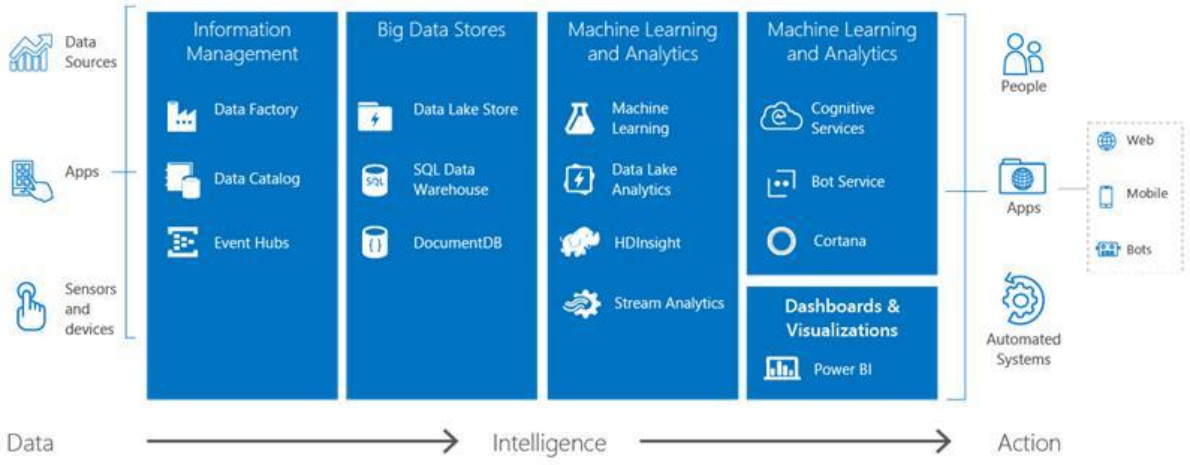
[702] Voir Hybrid Reward Architecture for Reinforcement Learning, juin 2017, qui décrit une architecture d'apprentissage par renforcement avec des agents fonctionnant en parallèle.

L'architecture d'IA de Microsoft est un maquis de briques logicielles avec un mix d'outils de développement, de déploiement dans le cloud et d'applicatifs horizontaux ou verticaux. Le « branding » de l'ensemble laisse à désirer et est encore plus instable que chez IBM qui a au moins « Watson » comme cri de raliement.

- L'**agent Cortana**, bon à tout faire, répond à la voix et joue le rôle de traducteur[703]. C'est un peu l'équivalent d'Amazon Alexa et de Google Assistant. Quelques bots ont été développés avec les APIs de Microsoft mais l'offre n'a rien à voir avec l'abondance qui sévit autour d'Amazon Alexa et plus récemment de Google Assistant.
- Les **services cognitifs** qui comprennent plus d'une vingtaine d'APIs sensorielles qui font de la reconnaissance d'images, du traitement du langage naturel (NLP), de la gestion de connaissances et de la recherche. A bas niveau, Microsoft propose en open source son framework de "deep learning" **CNTK** (Computational Network Toolkit) depuis fin 2015. Les API de vision artificielle permettent par exemple de détecter les émotions dans les visages et d'estimer l'âge des personnes (*ci-dessous*).

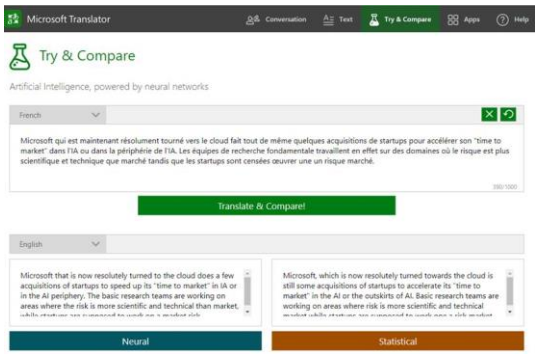
- Des **outils de création** d'applications avec l'IDE Visual Studio (Integrated Development Environment) et ses Code Tools for AI, et aussi l'outil Azure Machine Learning Studio, lancé en 2015, qui permet de créer ses modèles de machine learning et de les mettre en production. Annoncé en septembre 2017, Azure Machine Learning est maintenant découpé en trois modules avec Workbench pour la modélisation, Experimentation pour le test de modèles sur les infrastructures du cloud dont des GPU, et Model Management, pour les déploiements. Le tout avec une application native fonctionnant sous Windows et MacOS.

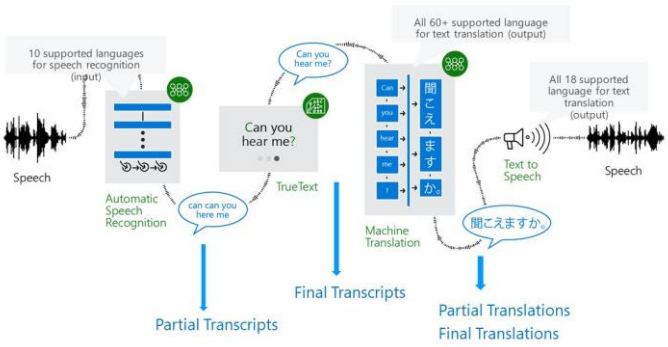
  La plateforme Azure est ouverte et intègre nombre d'outils open source du marché avec notamment TensorFlow et Caffe (frameworks), et aussi Apache Spark (pour la distribution des traitements sur les serveurs) et Docker (pour le déploiement d'applications).



[703] Voir A developer's guide to building AI applications – Create your first intelligent bot with Microsoft AI, 2018 (52 pages).

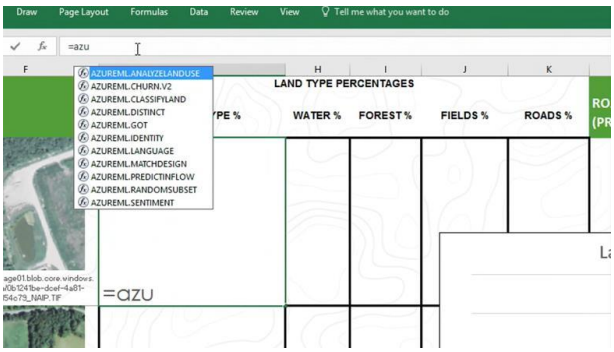
On trouve l'outil de création de chatbot **Microsoft LUIS** (Language Understanding Intelligent Service) lancé début 2017, l'outil de traduction **Microsoft Translator** et ses Translator Speech Translation API (*dont le processus est illustré ci-dessous*) ainsi que le **Microsoft Bot Framework** et le **BotBuilder** qui servent à créer son propre chatbot. Les **Visual Studio Tools** lancés sne 2018 permettent de gérer des projets avec les principaux frameworks du marché tels le CNTK de Microsoft, Tensorflow, Keras et Caffe.

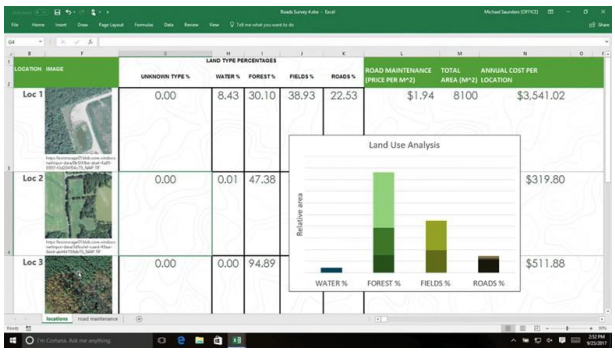




En mars 2018, Microsoft lançait **WinML**, un jeu d'APIs destinées à l'exécution sous Windows 10 de modèles de machine et deep learning déjà entraînés. Cela rappelle Core ML d'Apple qui est destiné à iOS. WinML doit intégrer la mise à jour de Windows 10 qui répond au doux nom de code Redstone 4. WinML est aussi calibré pour exploiter le chipset Intel Movidius qui est dédié à la reconnaissance d'images, notamment pour les webcams des laptops ou pour faire des recherches dans ses photos par une recherche naturelle sur leur contenu.

- L'infrastructure en cloud **Azure** et les logiciels serveurs de Microsoft et tous les outils de supervision qui vont avec.
- Dans l'embarqué, l'**Azure IoT Edge SDK** permet de gérer des modèles de deep learning entraînés de reconnaissance d'image dans des caméras utilisant le chipset mobile Qualcomm QCS603 et le Vision SDK de ce dernier. Microsoft va aussi intégrer **Windows ML** dans une mise à jour de Windows 10, un run-time permettant d'exécuter des modèles entraînés d'IA dans ses applications, en exploitant le CPU et le GPU de l'ordinateur.
- Diverses **applications** qui intègrent des briques d'IA, comme Office 365, Dynamics 365 et l'application mobile de gestion de photos Pix. Le traitement du langage à base d'IA est disséminé dans Office, et depuis des années. En septembre 2017, Microsoft annonçait l'intégration de services de machine learning du cloud Azure dans Excel, qui se manifestent sous la forme de fonctions (*ci-dessous*). **Seeing AI** est une application qui décrit l'environnement vu d'un smartphone, très utile pour les personnes mal-voyantes (vidéo).





Microsoft ajoutait des fonctionnalités d'IA dans **Dynamics**, les Dynamics 365 AI pour aider les commerciaux à prioriser leurs actions de prospection et de relance avec interrogation en langage naturelle, pour rendre les données clients accessibles via des chatbots et notamment via Cortana, puis « AI for Market Insights » qui « *vise à aider les équipes de marketing et de médias sociaux à accéder à des informations exploitables plus rapidement pour une réponse plus rapide aux besoins des clients* » ce qui ne veut strictement rien dire de précis. Il faut consulter le document Applications de gestion – Notes de publication Octobre 2018, juillet 2018 (273 pages) pour comprendre qu'il s'agit de tableaux de bords associés à l'analyse de sentiments dans les réseaux sociaux.

- On peut ajouter à tout cela **Microsoft Learn** qui est une offre de formation en ligne à tous ces outils. Et elle est gratuite. Vous pouvez par exemple vous former à la création d'un outil de classification automatique d'images[704].

Les données

Microsoft dispose d'une activité grand public, certes pas aussi soutenue que celle de Google, mais qui lui permet d'avoir une forte expertise dans le cloud ainsi que dans la captation de données d'usages, à même d'alimenter ses outils de machine et deep learning. Il en va ainsi du moteur de recherche **Bing**, de **Skype**, de **MSN**, de la console de jeu **Xbox** et bien évidemment de **Windows**.

Le matériel

Comme Google et IBM, Microsoft a développé sa propre architecture serveur pour gérer des réseaux de neurones. Elle s'appuie sur des processeurs développés en technologie FPGA. L'architecture s'inscrit dans le projet Brainwave dont les contours ont été dévoilés fin août 2017[705] et qui s'appuient sur :

- Une **architecture** de serveurs massivement parallèle et distribuée associant CPU et FPGA.
- De **processeurs FPGA** fabriqués par Intel en technologie 14 nm, les Stratix 10 (ex Altera)[706]. Ce sont des FPGA à mémoire qui stockent les paramètres des réseaux de neurones et évitent de faire appel à de la DRAM dans les serveurs, ce qui est bien plus rapide. Leur architecture de FPGA est dite « soft DNN » et donc reprogrammables, tandis que celle des TPU de Google n'est pas reprogrammable (« hard » DNN), ce qui apporte, pour faire simple, plus de flexibilité. L'architecture est optimisée à la fois pour des réseaux de neurones convolutionnels (Convnets, pour la reconnaissance d'images) qui nécessitent de multiplier des matrices et des réseaux de neurones récurrents (RNN, LSTM et consors, pour le traitement du langage). Cela leur apporte une plus grande flexibilité pour les déploiements à grande échelle dans leurs data-centers.
- Un **compilateur et un environnement d'exécution système** permettant de déployer des modèles du Microsoft Cognitive Toolkit tout comme de Google Tensorflow.

Cette architecture est déployée dans les datacenters de Microsoft Azure depuis 2016.

En parallèle, Microsoft Research planche aussi sur des ordinateurs quantiques topologiques à base de fermions de Majorana. Je décris leur approche dans l'ebook Comprendre l'informatique quantique, septembre/novembre 2018 (342 pages).

---

[704] Voir Classifier des images avec le Service Vision personnalisée de Microsoft.

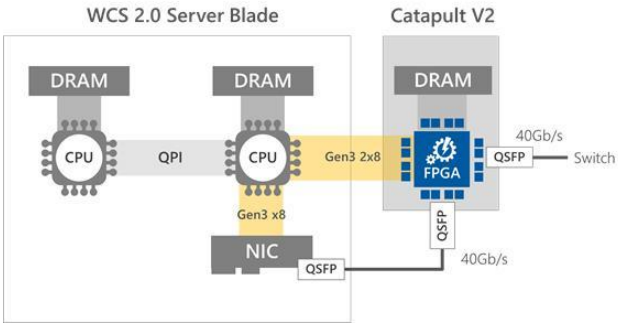
[705] Voir Microsoft unveils Project Brainwave for real-time AI, août 2017.

[706] Microsoft avait produit ses premiers FPGA en 2011 dans ses serveurs Catapult V0 pour la gestion d'index de Bing. La V1 de Catapult sortait en 2012. En 2013, 1600 FPGA étaient mis en production. Capatupt V2 sortait en 2014 avec une architecture de bus optimisée pour faire communiquer les CPU avec les FPGA dans les serveurs, via un bus PCI à 64 Gbits/s (4 canaux).

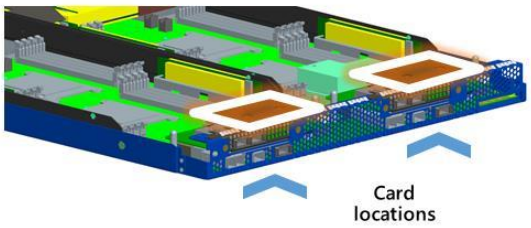
Intel/Altera Stratix 10

- 10 TFLOPS FP32
- HBM2 integrated
- Up to 1 GHz
- 14nm process
- 80 GFLOPS/W

Catapult v2 Mezzanine card



WCS Gen4.1 Blade with NIC and Catapult FPGA



Facebook

Le leader occidental des réseaux sociaux est avide d'IA à tous les étages pour améliorer l'ensemble de ses services que ce soit pour caser les bons contenus et les bonnes publicités aux bons endroits.

Les logiciels

Facebook est à l'origine de framerworks de machine learning et deep learning : PyTorch, Caffe2 et ONNX.

Leurs solutions logicielles d'IA vont des fonctions de reconnaissance de photos dans ses différents services mobiles à tout ce qui permet de mieux cibler les publicités pour optimiser les revenus en passant par le **Bot Framework** lancé en avril 2016 servant à la création de chatbots s'intégrant dans l'application Facebook Messenger.

facebook

offre	forces	faiblesses
API chatbots pour Messenger traitement d'images dans Facebook et Instagram framework de machine learning PyTorch	services grand public données utilisateurs Instagram et Whatsapp Instant Messenger chatbot laboratoire de recherche à Paris / Montréal / Yann Le Cun – Jérôme Pesenti	mal équipé pour accompagner les grandes entreprises pas de véritable plateforme logicielle pour les entreprises image suite à l'affaire Cambridge Analytica

La société a plus d'une centaine de chercheurs en IA dont le célébrissime Yann LeCun, inventeur des réseaux de neurones convolutionnels qui sont la base de la reconnaissance d'images dans le deep learning. Il gère l'équipe de chercheurs du FAIR (Facebook Artifical Intelligence Research) qui sont installé à Paris (dirigée par Antoine Bordes), Montréal (dirigée par Joëlle Pineau), Menlo Park en Californie, New York et Seattle.

Yann Lecun est assez remonté contre les visions anxyogènes de l'IA[707]. Yann LeCun est aussi professeur au Collège de France sur le deep learning[708]. Il reporte depuis début 2018 à Jérome Pesenti, un autre Français de l'IA passé par IBM Watson et CEO de **Benevolent.ai** (2013, UK, $208M) et patron de Facebook AI.

---

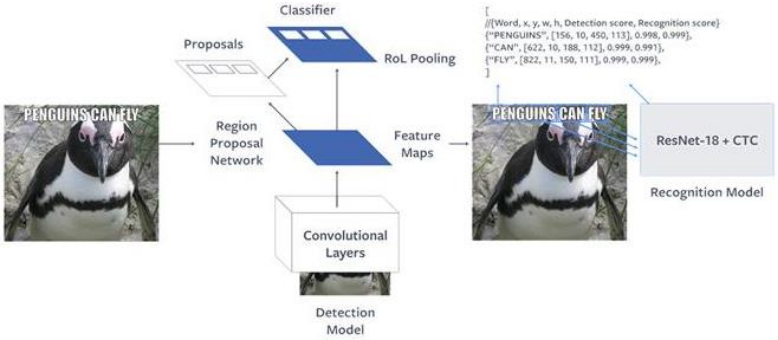
[707] Voir Facebook's head of AI wants us to stop using the Terminator to talk about AI, octobre 2017.

[708] Voir sa leçon inaugurale de février 2016 : L'apprentissage profond : une révolution en intelligence artificielle, qui fait un très bon panorama technique du machine learning et du deep learning, ainsi que les autres sessions dans http://www.college-de-france.fr/site/yann-lecun/_audiovideos.htm.

Parmi les projets de recherche de Facebook, on compte notamment **DeepFace**, une solution de deep learning de reconnaissance des visages avec un réseau de neurones de neufs couches et 120 million connexions entraîné sur quatre millions d'images. La précision du système serait de 97%. Le système a bénéficié de la contribution de Yaniv Taigman, issu de **Face.com** (2005, Israël, $11M), acquise en 2007.

Facebook veut notamment utiliser la reconnaissance d'images pour informer ses utilisateurs malvoyants du contenu visuel de leur timeline.

D'autres projets de recherche marquants comprennent **Rosetta** qui reconnaît les textes dans les images et vidéos[709] et des outils pour aider les développeurs à détecter **(SapFix)** et corriger **(Sapienz)** les bugs dans leur code[710]. Reste à en apprécier la portée exacte !



Les autres acquisitions de Facebook dans l'IA comprennent :

- **Jibbigo** (2009, USA) en 2013, pour son système de traduction speech-to-speech.
- **Pebbles Interfaces** (2010, Israël, $14,45M) en 2015, pour son système de captation de gestes.
- **wit.ai** (2013, USA, $3M) en 2015, une petite startup de Palo Alto, pour ajouter des fonctionnalités de reconnaissance de la parole dans ses services et notamment de Messenger. Mais Wit.ai est aussi une plateforme utilisée par des milliers de développeurs. La startup avait été créée par trois français : Alexandre Lebrun, Willy Blandin et Laurent Landowski.
- **Surreal Vision** (2014, UK), une startup acquise par Oculus en 2015, spécialisée dans vision 3D.
- **Faciometrics** (2015, USA) en 2016, pour sa solution mobile d'analyse de visages.
- **Ozlo** (2014, USA, $14M) fin juillet 2017, qui aide à trouver un bon restaurant.
- **Blomsbury** (2015, USA, $1,7M) en juillet 2018, qui permet d'interroger en langage naturel des documents non structurés. C'était l'objet de leur projet Cape, publié en open source fin 2017. L'outil doit en théorie permettre de lutter contre la propagation de fake news.

Les données

Facebook est un gros collecteur de données générées par les utilisateurs de ses services.

- Le **texte** via Facebook et Whatsapp.
- Les **photos** via Facebook et Instagram.
- Les **vidéos** via Facebook.

[709] Voir Rosetta: Understanding text in images and videos with machine learning, septembre 2018.

[710] Voir Finding and fixing software bugs automatically with SapFix and Sapienz, septembre 2018.

En règle générale, Facebook ne vend pas de données mais l'accès à des utilisateurs en fonction de profiling issus de ces données. Mais l'affaire Cambridge Analytica au printemps 2018 a montré que le système avait des trous dans la passoire, les éditeurs d'applications pouvant dans certaines conditions récupérer des informations de profiling des utilisateurs. Depuis, Facebook a fait le ménage dans ses APIs mais rien n'est sûr à 100%.

Le matériel

Enfin, en octobre 2018, Facebook faisait sa dernière incartade dans le matériel[711] avec le lancement de **Portal**, un écran pour les communications vidéo suportant aussi Amazon Alexa. Un cache est fournit pour sa caméra pour préserver la vie privée. Facebook insiste beaucoup sur le « privacy by design » intégré dans l'objet pour protéger la vie privée des utilisateurs. Ce qui n'a pas empêché les échos négatifs dans une partie des médias, surtout aux USA[712].

Sinon, Facebook utilise aussi le machine learning dans la gestion de ses data centers, comme décrit dans Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective, 2017 (11 pages).

[711] Après, bien entendu, son acquisition d'**Oculus Rift** (2012, USA, $96M) en 2014.

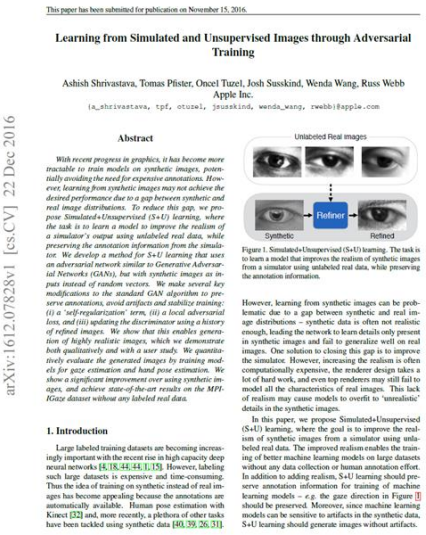
[712] Voir Facebook's launch of Portal has been stymied by trust issues de Casey Newton, octobre 2018.

Apple

Apple est bien plus orienté produits et marchés que ne le sont IBM et Microsoft. Non seulement la société n'a pas formellement de laboratoire de recherche fondamentale mais elle ne publiait jusqu'en 2016 *aucun* papier dans le domaine de l'IA. C'est tout le contraire de l'innovation ouverte !

Elle est cependant sortie du bois en publiant un premier papier fin 2016 sur la reconnaissance d'images[713]. La suite se trouve dans leur Apple Machine Learning Journal.

L'IA est maintenant incontournable dans l'offre d'Apple, tant au niveau logiciel (SIRI, photos) que matériel (chipsets A11 et A12 avec NPU, HomePod, incartades dans les véhicules autonomes).



This paper has been submitted for publication on November 15, 2016.

Learning from Simulated and Unsupervised Images through Adversarial Training

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, Russ Webb
Apple Inc.
{a_shrivastava, tpf, otuzel, jsusskind, wenda_wang, rwebb}@apple.com

arXiv:1612.07828v1 [cs.CV] 22 Dec 2016

Abstract

With recent progress in graphics, it has become more tractable to train models on synthetic images, potentially avoiding the need for expensive annotations. However, learning from synthetic images may not achieve the desired performance due to a gap between synthetic and real image distributions. To reduce this gap, we propose Simulated+Unsupervised (S+U) learning, where the task is to learn a model to improve the realism of a simulator's output using unlabeled real data, while preserving the annotation information from the simulator. We develop a method for S+U learning that uses an adversarial network similar to Generative Adversarial Networks (GANs), but with synthetic images as inputs instead of random vectors. We make several key modifications to the standard GAN algorithm to preserve annotations, avoid artifacts and stabilize training: (i) a 'self-regularization' term, (ii) a local adversarial loss, and (iii) updating the discriminator using a history of refined images. We show that this enables generation of highly realistic images, which we demonstrate both qualitatively and with a user study. We quantitatively evaluate the generated images by training models for gaze estimation and hand pose estimation. We show a significant improvement over using synthetic images, and achieve state-of-the-art results on the MPIIGaze dataset without any labeled real data.

1. Introduction

Large labeled training datasets are becoming increasingly important with the recent rise in high capacity deep neural networks [4, 18, 44, 44, 1, 15]. However, labeling such large datasets is expensive and time-consuming. Thus the idea of training on synthetic instead of real images has become appealing because the annotations are automatically available. Human pose estimation with Kinect [32] and, more recently, a plethora of other tasks have been tackled using synthetic data [40, 39, 26, 31].

Unlabeled Real images

Refiner

Synthetic

Refined

Figure 1. Simulated+Unsupervised (S+U) learning. The task is to learn a model that improves the realism of synthetic images from a simulator using unlabeled real data, while preserving the annotation information.

However, learning from synthetic images can be problematic due to a gap between synthetic and real image distributions – synthetic data is often not realistic enough, leading the network to learn details only present in synthetic images and fail to generalize well on real images. One solution to closing this gap is to improve the simulator. However, increasing the realism is often computationally expensive, the renderer design takes a lot of hard work, and even top renderers may still fail to model all the characteristics of real images. This lack of realism may cause models to overfit to 'unrealistic' details in the synthetic images.

In this paper, we propose Simulated+Unsupervised (S+U) learning, where the goal is to improve the realism of synthetic images from a simulator using unlabeled real data. The improved realism enables the training of better machine learning models on large datasets without any data collection or human annotation effort. In addition to adding realism, S+U learning should preserve annotation information for training of machine learning models – *e.g.* the gaze direction in Figure 1 should be preserved. Moreover, since machine learning models can be sensitive to artifacts in the synthetic data, S+U learning should generate images without artifacts.

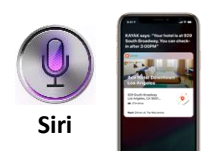
Les logiciels

Les acquisitions d'Apple sont peu nombreuses en règle général même si elles eu tendance à s'accélérer depuis le décès de Steve Jobs en 2011.

Dans l'IA, on peut compter **Turi** (2013, USA, $25M) en 2016 pour $200M qui avait développé un projet d'analytics à base de machine learning, **Emotient** (2006, USA, $6M) en 2016 pour la reconnaissance des visages et des émotions, **VocalIQ** (2011, UK, $1,2M) en 2015 qui devait enrichir les fonctionnalités de reconnaissance de la parole de SIRI en ajoutant de l'auto-apprentissage, ainsi que **Perceptio** (2014, USA) en 2015, dans la reconnaissance d'images s'appuyant sur du deep learning et **Regaind** (2015, France, 400K€) en septembre 2017.

Ont suivi les acquisitions de **Lattice Data** (2015, USA) en 2017, issu du projet DeepDive de Stanford, spécialisé dans l'extraction de données de contenus non structurés, et **SensoMotoric Instruments** (1991, Allemagne) qui a développé une technologie de suivi du regard à base d'IA (eye tracking).

SIRI est de son côté le résultat de l'acquisition en 2010 de la startup **SIRI** (2007, USA, $24M) en 2010, elle-même issue d'un projet de **SRI International** financé par la DARPA, et de l'usage des technologies issues de l'américain **Nuance Communications**, la société leader du secteur de la reconnaissance de la parole qui fait plus de $2B de chiffre d'affaire !



Siri

Homepod

Ce dernier utilise en partie des technologies issues de Scansoft, provenant du belge Lernout & Hauspie qui avait acquis la technologie de reconnaissance de la parole de Ray Kurzweil !

Apple sinon comble les trous dans son offre d'IA via son partenariat avec IBM qui porte notamment sur Watson, une manière indirecte de séduire les grandes entreprises et les DSI, qui étaient les grandes bêtes noires de Steve Jobs.

---

[713] Voir Learning from Simulated and Unsupervised Images through Adversarial Training en décembre 2016.

Il est cependant probable qu'Apple devra faire quelques acquisitions dans le cadre de son projet de voiture autonome.

Apple lançait son framework **CoreML** sur iOS en juin 2017. C'est un framework qui permet l'exécution de modèles d'IA déjà entraînés sur smartphones. La version 2 de Core ML a été lancée un an plus tard. C'est un framework généraliste qui permet de gérer des modèles entraînés de deep learning aussi bien que de machine learning classique avec des régressions linéaires, non linéaires, du SVM et autres modèles probabilistes du genre.

Les données

Apple gère de gros volumes de données de ses utilisateurs, que cela soit via iCloud ou son magasin d'applications mobiles qui lui permet de connaître pas mal de choses sur leurs habitudes. Ceci étant, ces données ne sont pas utilisées par Apple avec la même logique qu'un Google ou un Facebook. Apple se positionne même explicitement comme étant plus respectueux de la vie privée de ses utilisateurs.

Derrière les bonnes intentions, l'origine en est limpide : c'est lié à leur modèle économique de vente de matériel qui domine leur mix de revenu alors que Google et Facebook dépendent presque exclusivement du revenu publicitaire.

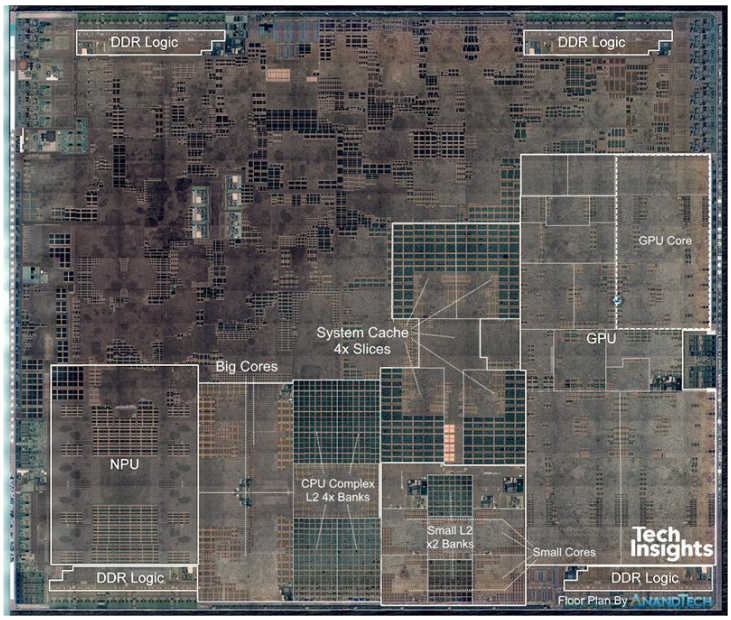
Le matériel

Apple utilise beaucoup d'IA dans ses iPhone. Les iPhone 8 et X annoncés en septembre 2017 intègrent le chipset A11 Bionic et sa fonctionnalité neuromorphique Neural Engine, dédiée à l'exécution d'applications de deep learning, comme pour la fonction FaceID de reconnaissance de visage ainsi que pour la reconnaissance de la parole avec SIRI.

Il y a fort à parier que cela permettra de faire émerger de nombreuses applications mobiles exploitant cette capacité.

La série d'iPhones Xs lancée en septembre 2018 intègre le chipset A12 réalisé en technologie 7 nm avec un doublement des capacités de son NPU (*ci-contre*).

Le NPU est le rectangle en bas à gauce du processeur[714]. Il a une capacité d'environ 4 Tops (téra opérations par seconde, en entier). Il est principalement exploité par les fonctions photos, de login par la détection du visage et de reconnaissance vocale des smartphones.



Apple a sinon lancé ses enceintes à commande vocale équipées de SIRI, les Homepod en 2018. Etant dans le troisième écosystème du marché des enceintes vocales après les Amazon Echo et celles qui supportent Google Assistant, leur marché est surtout celui des Apple afficionados.

Enfin, Apple investi dans le domaine des véhicules autonomes. On a commencé à en entendre parler presque officiellement avec le projet Titan.

---

[714] Source de l'image : The iPhone XS & XS Max Review: Unveiling the Silicon Secrets de Andrei Frumusanu, octobre 2018.

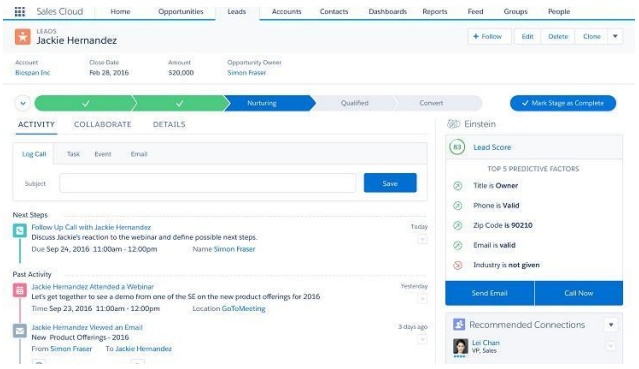
Il semblerait que leur approche dans un premier soit celle de l'équipementier, créant des systèmes de conduite autonomes avec matériel et logiciel, destinés à des constructeurs automobiles. Cela fait longtemps que des rumeurs circulent sur leur vélléité de construire des véhicules, ou de faire l'acquisition de Tesla. Mais elles n'ont pas été confirmées à ce jour.

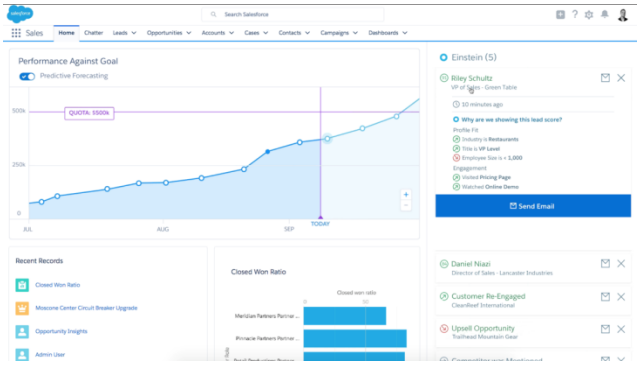
Salesforce

Chez Salesforce, l'offre d'IA s'appelle modestement **Einstein**, une offre d'IA en cloud au service des forces de vente.

L'offre qui s'appelle précisément **Einstein High Velocity Sales Cloud** comprend les briques suivantes :

- **Einstein Lead Scoring** : avec des outils à base de machine et deep learning de repérage des meilleurs leads d'un pipe commercial en fonction d'une analyse multicritères (pour peu que la base soit bien renseignée…). Un cas d'usage classique du machine learning en tout cas.
- **Einstein Activity Capture** : capture des informations utiles dans les mails et calendriers et des modèles personnalisés de réponses par e-mails.
- **Lightning Sales Console** : un espace de travail personnalisable de suivi des meilleurs leads.
- **Lightning Dialer** : pour contacter les prospects en un clic.
- **Salesforce Engage** : notifications en temps réel d'opportunité d'interaction client.
- **Salesforce AppExchange** : un écosystème de solutions tierces-parties.





- **Einstein Bots**, sa plateforme de création de chatbots lancée en novembre 2017 qui exploite **Einstein Language**, l'outil de traitement du langage de Salesforce. Le système détecte le niveau d'émotion de l'utilisateur (positif, négatif ou neutre). Einstein Bots comprend une interface graphique pour la conception de chabots. Ces chatbots sont destinés aux clients de l'entreprise exploitant Salesforce tout comme à leurs commerciaux.
- **Einstein Prediction Builder**, un système de création de prévisions applicable à toute donnée gérée par Salesforce. Et pour ceux qui veulent développer, Salesforce faisait un don à la communauté open source qui demande le respect avec **AutoML TransmogrifAI** qui comprend 22 lignes de code, commentaires compris. Elle s'appuie sur Apache Spark et est écrite en Scala[715].

La recherche en IA chez Salesforce s'appuie sur 100 ingénieurs pour la recherche fondamentale et 200 ingénieurs pour la recherche appliquée.

[715] Ce n'est pas une blague. Le code est ici : https://transmogrif.ai/.

Une bonne partie d'Einstein provient de briques et de compétences récupérées par SalesForce à l'occasion de l'acquisition de diverses startups : **RelateIQ** (2011, USA, $69M) en 2014, qui était spécialisées dans la relation client, devenu SalesForceIQ, une version de SalesForce pour les PME, **PredictionIO** (2013, USA, $2,7M) acquis en 2016 pour sa solution open source de machine learning et **MetaMind** (2014, USA, $8M) acquis en 2016 pour ses solutions de reconnaissance d'image, soit l'équivalent de 175 data scientists. Einstein s'appuie aussi sur les APIs d'**IBM Watson**.

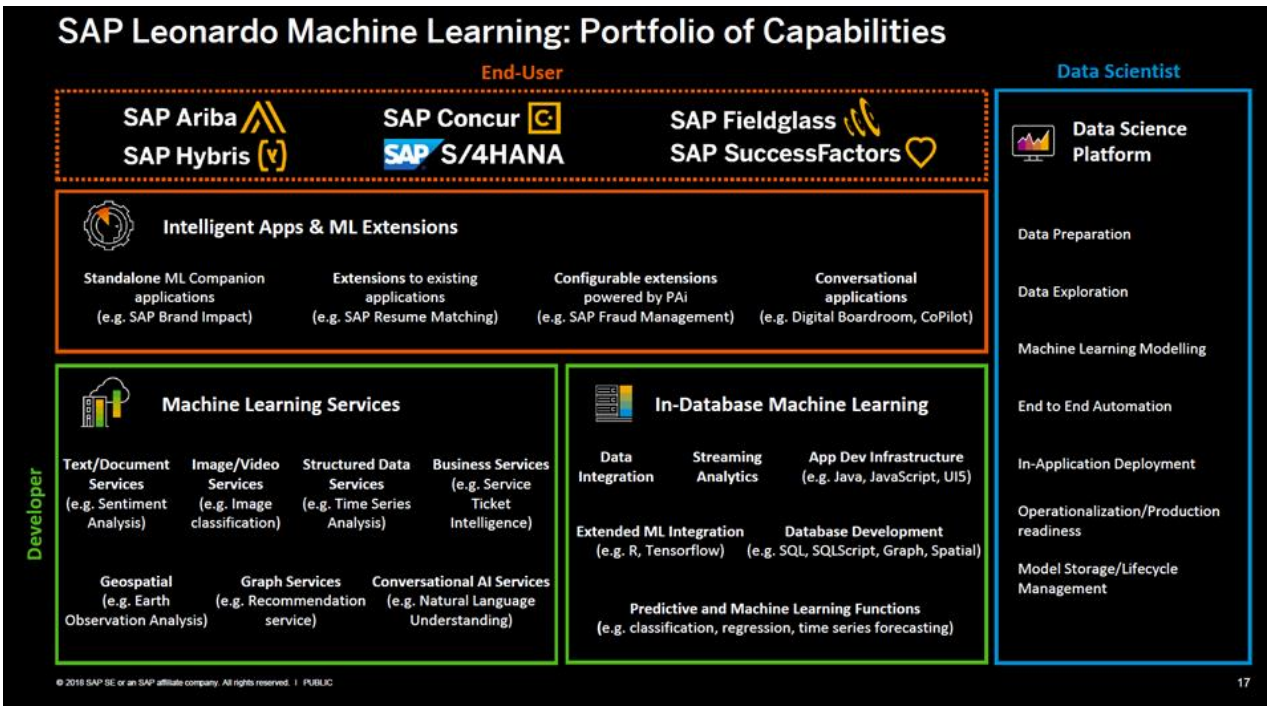
SAP

L'IA de SAP est intégrée depuis l'été 2017 dans le vaste gloubi-boulga qu'est la plateforme **Leonardo**. Les équipes marketing de SAP méritent largement d'aller dans l'enfer des marketeurs tellement les évolutions de leur offre en IA sont difficiles à suivre avec des dizaines de briques difficiles à positionner les unes par rapport aux autres.

Les briques logicielles de Leonardo comprennent l'ensemble des outils servant à exploiter les données générées par les applications métiers transactionnelles de SAP. Leonardo s'appuie sur divers outils de machine learning autour de **SAP Leonardo Machine Learning Foundation** (vidéo) qui couvre tous les besoins de gestion de données, de textes et d'images. La plateforme comprend aussi un moteur de gestion de règles (Rules Engine). Cette plateforme de machine learning s'appellait précédemment Clea.

Côté outils de travail, on y trouve notamment l'APL (**Automated Predictive Library**) pour les analystes et la PAL (**Predictive Analysis Library**) qui est destinée aux data scientists.

Côté couches basses, Leonardo supporte aussi l'intégration de modules développés en **TensorFlow**. SAP fait la promotion du « *in-database machine learning* », même si ce niveau d'intégration n'est pas bien précisé ni à quelle base de donnée cela correspond. Cela semble en fait relever du « *in-memory processing* ». Il s'agit en tout cas de machine learning s'appuyant sur des méthodes classiques : arbres de décision, SVM, K-means, régressions linéaires et non linéaires, etc. Les modèles plus complexes avec des réseaux de neurones et du deep learning sont mis en œuvre avec Tensorflow.



SAP décline toutes ces plateformes dans de nombreux cas d'usage et marché verticaux et en particulier dans l'industrie et les objets connectés.

L'éditeur exploite même des outils de reconnaissance d'images pour certains besoins, tels **SAP Brand Impact** qui mesure la visibilité de marques dans les médias ou dans un stade ou **SAP Resume Matching** pour la recherche de correspondance de CV et de postes ouverts dans la RH.

Enfin, côté serveurs, SAP est partenaire à la fois d'**Intel** et **Nvidia** pour l'exécution des briques de machine learning de Leonardo, en particulier sur les serveurs DGX de Nvidia.

Oracle

Avec SAP, Oracle est l'un des plus gros éditeurs de logiciels d'entreprise mondiaux. Il était plutôt discret dans son intégration de l'IA dans la stratégie jusqu'en 2017. Ils se sont embarqués dans l'IA en intégrant progressivement et ouvertement divers outils de machine learning dans leurs logiciels d'infrastructures horizontaux et applications métiers verticales[716]. A commencer par la quinquaillerie permettant de créer des chatbots, le type de projet d'IA le plus courant dans les entreprises.

On en trouve ainsi dans **Oracle Management Cloud Services**, dans **Oracle Advanced Analytics, Oracle Data Miner** et **Oracle Internet of Things Cloud Service**, qui sont des outils d'analytics divers, exploitant des arbres de décision et générant des rapports divers.

Oracle annonçait son **Chatbots for Developers** en 2016 et une plateforme de développement associée et le service associé, **Intelligent Bot Cloud Service** (IBCS). Leur outil permet de créer des chatbot à commande textuelle ou vocale. Je n'ai pas trouvé son origine. Il fait évidemment appels aux outils de gestion de bases de données et de cloud d'Oracle.

En octobre 2018, Oracle mettait à jour son offre de chatbot avec une solution d'assistant numérique intégrant l'accès aux applications d'ERP, de CRM ainsi que de RH. Cet Oracle Digital Assistant est commandable par la voix (via Amazon Alexa) ou par le texte (via Slack, Facebook Messenger, WeChat) (vidéo).

Quid des acquisitions dans l'IA ? Oracle est assez friand de startups, mais curieusement, pas vraiment dans l'IA. On note surtout celle de **Crosswire** (2013, Israël, $5M) en 2016 qui propose une solution cross-device de ciblage et d'analytics publicitaires doté d'outils de présentation graphique. Et puis celle de **Datascience** (2014, USA, $28M) en juin 2018, un spécialiste de la data science et du machine learning.

Startups

D'un point de vue technique, nous avons bien vu que l'IA se mettait en œuvre avec un ensemble de techniques assez disparates, presque toutes disponibles en open source, assez concentrées sur le traitement du langage, de l'image et des données et avec une gradation assez forte dans le niveau d'intelligence des solutions. Nous avons aussi observé la forte dépendance entre les solutions d'IA et les données qui les alimentent.

[716] Accenture et Oracle publiaient un livre blanc décrivant la stratégie IA d'Oracle, Technology Vision for Orange Executive Summary (32 pages). Ca parle d'entreprise intelligence et d'IA humaine. Mais franchement, pas beaucoup d'IA !

Caractéristiques

L'IA est devenue un phénomène de mode pour les startups comme l'ont été les réseaux sociaux (vers 2004), la vidéo (vers 2006), la mobilité (à partir de 2009), les objets connectés (vers 2011), le cloud ou la Blockchain (depuis 2017). Une startup qui ne brandit pas l'IA comme sauce magique parait dépassée par les événements. Et nous avons déjà des sous-modes avec les chatbots, la robotique, la cognitif, etc.

On sourit souvent des startups qui ont "créé une IA" alors qu'elles ont correctement paramétré un réseau de neurones avec TensorFlow à partir d'exemples et après avoir tâtonné, ou qu'elles ont utilisé une vieille technique de prévision ou de clustering à base de machine learning.

Ces effets de mode sont notamment alimentés par les prévisions de chiffre d'affaires des analystes, comme IDC qui prédit que le marché de l'IA représentera $46B de CA en 2020[717]. Ces évaluations ont toujours tendance à gonfler des chiffres qui sont incalculables. Ainsi, quel est le CA en IA de Google, Facebook ou Microsoft, qui ne sont d'ailleurs visiblement pas intégrés dans l'estimation d'IDC ? Une donnée intéressante serait d'évaluer le CA additionnel généré par l'IA chez les éditeurs de logiciels et startups, mais la part de l'IA dans la valeur ajoutée d'un Oracle ou d'un SalesForce est bien difficile à évaluer. On a connu le même phénomène avec les prévisions sur le marché des objets connectés[718].

Cela amène la généralisation du phénomène de l'*IA washing*, décrivant ces startups qui usent et abusent de la terminologie de l'IA, souvent, sans préciser la manière dont leur solution en tire parti. Ce qui ne veut pas dire qu'elles ne font pas d'IA pour autant. Elles le font à des degrés très variables.

Depuis début 2016, ma position sur la question a évolué. Les techniques de l'IA sont largement disponibles, en open source, dans le cloud et dans l'embarqué. Les techniques du machine learning sont relativement faciles à mettre en œuvre. Celles du deep learning nécessitent un effort conceptuel plus important, mais accessible aux jeunes développeurs et data scientists sans compter les boîtes à outils prêtes à l'emploi comme chez Clarifai pour l'analyse d'images.

L'IA est en train de devenir l'équivalent moderne du développement web : un ensemble de techniques de plus en plus abordables.

Si une entreprise indique avoir « *créé une IA* », il faut traduire par « *nous avons créé une solution logicielle et/ou matérielle qui intègre des briques technologiques de l'IA et souvent d'autres briques techniques* ». Pour d'autres, une IA est un système antropomorphique qui imite une caractéristique humaine, comme la compréhension du langage. Pour moi, c'est un logiciel qui intègre une ou plusieurs briques logicielles du vaste champ scientifique de l'IA depuis sa création en 1955. Cela intègre le machine learning, le deep learning, les moteurs de règles tout comme les réseaux multi-agents. Et une IA logicielle n'est pas forcément antropomorphique. Elle peut réaliser des opérations qui sont inaccessibles à l'intelligence et à la mémoire humaines. Un système de segmentation automatique d'une énorme base de données réalise une tâche sur-humaine !

C'en est au point où comme l'Internet ou la mobilité, l'IA est en train de devenir incontournable. Progressivement, l'ensemble des solutions logicielles du marché vont intégrer des briques d'IA. N'importe quel logiciel de gestion ou logiciel métier, par exemple, utilisera le machine learning pour présenter des analyses des données générées et faire des prévisions. Tout logiciel exploitant du texte intègrera des fonctions de traitement du langage. Tout logiciel exploitant des images va s'appuyer sur du deep learning pour déterminer ce que contiennent les images.

---

[717] Voir Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $12.5 Billion This Year, According to New IDC Spending Guide, avril 2017.

[718] Que j'avais eu l'occasion de décrire dans La grande intox des objets connectés en août 2015.

Laurent Alexandre aime à dire que les investissements dans les startups de l'IA ne valent pas grand-chose et que leurs investisseurs sont des gogos, les seuls créateurs de véritables IA étant les GAFA. C'est évidemment exagéré ! Nombre de startups créent des IA aussi sophistiquées que celles des GAFA. En fait, le risque est simplement que l'usage de l'IA dans les logiciels devienne une simple commodité, donc une banalité. Comme le fait de créer un logiciel sous la forme d'un couple entre serveur web et application exploitable à partir d'un navigateur. D'ici 5 à 10 ans, le marché de l'IA se confondra donc quasiment avec l'ensemble du marché du logiciel et d'Internet.

Les GAFA créent et utilisent des IA sous la forme de logiciels open source et de data centers. Les algorithmes innovants sont encore créés en masse par des chercheurs issus d'universités du monde entier. Les GAFA n'ont pas encore le monopole de la créativité dans la recherche. Qui plus est, les solutions techniques sont open source et ne doivent pas obligatoirement tourner dans les infrastructures des GAFA. Comme l'Internet, l'IA est tout de même très distribuée.

Derrière l'habillage marketing, il reste à comprendre ce que la startup a réellement produit : a-t-elle assemblé des briques logicielles existantes de manière traditionnelle, a-t-elle créé des briques spécifiques, a-t-elle juste entraîné un modèle assez simple et mis en forme des données, d'où viennent-elles et la solution est-elle une simple application directe de techniques existantes ? En général, c'est bien le cas. Mais le choix, la programmation et l'entraînement d'un modèle de deep learning ou de machine learning pour répondre à un besoin spécifique requiert des compétences encore rares.

On ne devient pas développeur dans l'IA du jour au lendemain, de même qu'il a fallu du temps pour que les développeurs d'applications procédurales ou client serveur s'adaptent à la programmation événementielle du web et avec ses nombreux framework qui changent tout le temps (jquery, Angular, React, Node). Selon IDC, 1% des logiciels utiliseraient de l'IA auourd'hui et en 2018, 75% des développeurs intègreront de l'IA dans leur code, ce qui est probablement un peu optimiste, ne serait-ce que pour tenir compte du laps de temps pour se former[719].

L'IA en est encore au stade artisanal et du bricolage. Cela ne se voit évidemment pas directement quand on fait le tour d'horizon des startups du secteur. Surtout dans le mesure où la plupart d'entre elles sont "b-to-b" et diffusent leurs solution en marque blanche. Vous les retrouverez éventuellement dans les agents conversationnels des sites web de marques, dans le ciblage marketing qui vous touche avec une offre pertinente (ou pas du tout…), dans des robots capables de dialoguer plus ou moins avec vous, ou dans les aides à la conduite de votre voiture haut de gamme.

L'un des moyens de se rendre compte indirectement de cet aspect artisanal consiste à d'évaluer la part produit et la part service des entreprises du secteur. Plus la part du produit est faible, plus on est dans le domaine de l'artisanat. Cela n'apparaît pas dans les données publiques mais peut au moins d'obtenir quand on a l'occasion d'observer à la loupe ces entreprises : dans le cadre d'une relation grand compte/startup, d'un investissement ou même d'un recrutement. On peut l'observer également dans les profils LinkedIn des salariés de l'entreprise s'ils sont disponibles. Bref en utilisant ce que l'on appelle des sources d'information "ouvertes".

Les startups de l'IA, surtout américaines, ont quelques points communs marquants :

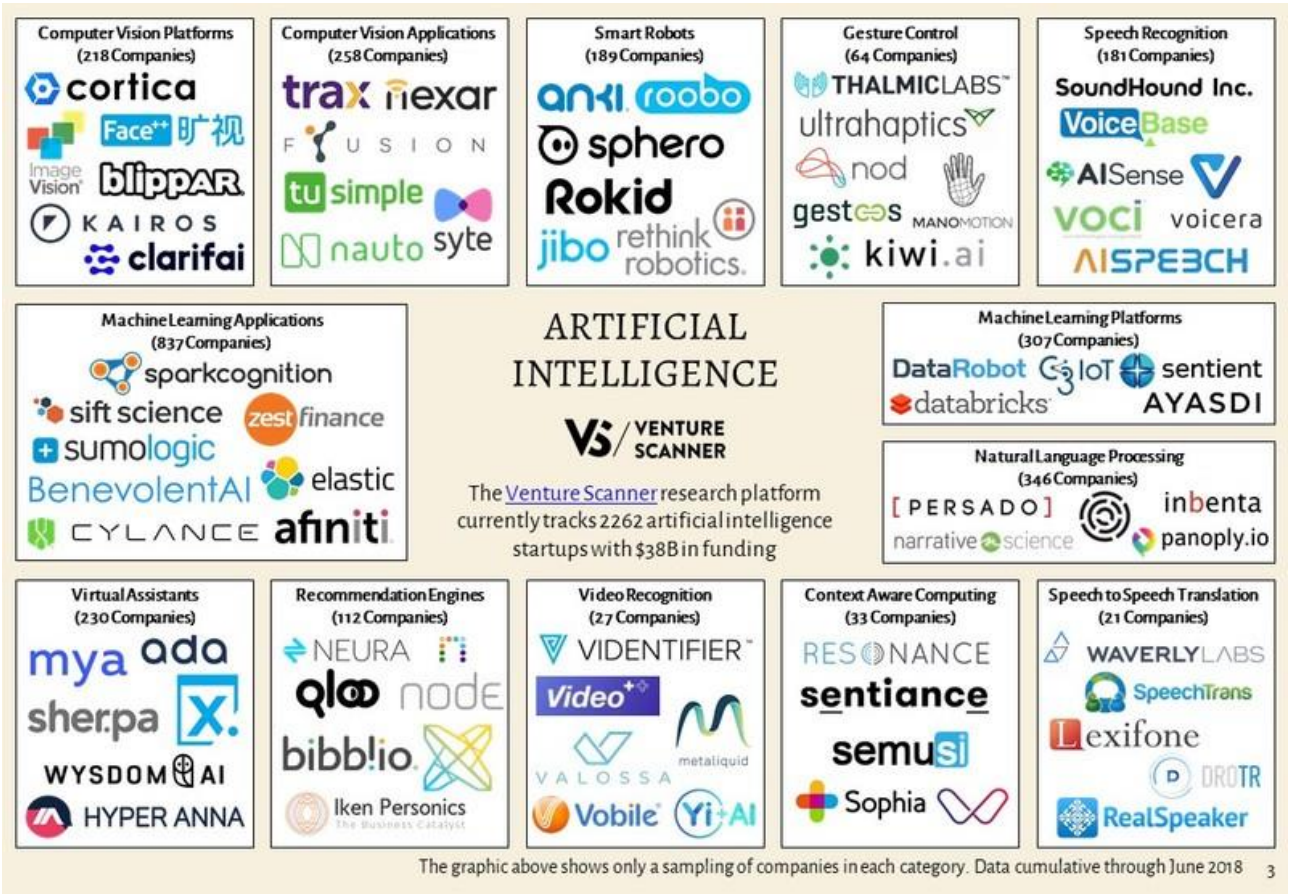
- Elles ont majoritairement des **approches marché "b-to-b"** avec des marchés visés qui sont toujours les mêmes, entre horizontal et vertical. Exemples de marché sursaturés : la détection de fraudes dans la finance et et l'analyse prédictive du comportement des consommateurs dans le marketing en ligne et mobile.
- Aux USA, on y trouve souvent les ombres de la **DARPA**, de la **NSA** et de la **CIA** comme clients voire même comme investisseurs pour cette dernière, via son fonds **InQTel**. Surtout pour les solutions "horizontales". Ce n'est pas une question de "Small Business Act" mais simplement de besoins de ces organisations de défense et de renseignement !

[719] Source : IDC FutureScape: Worldwide IT Industry 2017 Predictions.

- On retrouve aussi beaucoup d'anciens des Universités de **Stanford** et du **MIT** dans les startups de l'IA, généralement bardés d'un ou de plusieurs PhD en IA. Ils sont issus du monde entier.
- Les **technologies d'IA** employées sont assez mal documentées. Le machine learning et le deep learning reviennent souvent sans que l'on puisse évaluer si les startups ont réellement fait avancer l'état de l'art. Comme il se doit, une startup doit présenter un risque marché plus qu'un risque technologique ou scientifique. C'est pourquoi les startups de l'IA sont généralement positionnées dans l'application de techniques d'IA connues à des marchés divers, horizontaux ou verticaux. Elles profitent aussi parfois de l'effet d'opportunité en labellisant "IA" des projets qui quelques années auparavant auraient été vendus sous le sceau du "big data".
- Les solutions sont très souvent proposées sous la forme **d'APIs en cloud** mais les approches plateformes sont encore émergentes car elles ne bénéficient pas d'un effet push/pull courant dans le grand public (la demande pour des smartphones Android entrainant celles d'applications tournant dessus).
- Les **levées de fonds** sont encore relativement modestes dans l'ensemble. On dépasse dans de rares cas les $100M. Ce n'est pas beaucoup par rapport à plus de $1B réalisées par des licornes telles que Pinterest ou MagicLeap. Les licornes sont d'ailleurs presque toutes des startups grand public.

Cartographie

Il existe de nombreuses sources de cartographies de startups de l'intelligence artificielle, notamment chez **CBInsights**, **VentureScanner** et **FranceIsAI**, pour le marché français dans ce dernier cas. Ces cartographies sont apparues vers 2015 lorsque l'investissement dans l'IA a commencé à décoler.

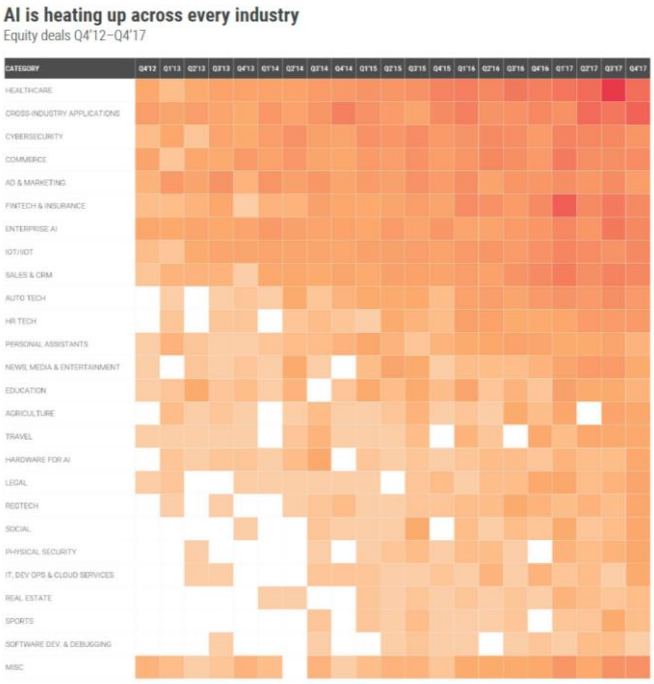


Je vais m'appuyer sur ce suivi du secteur par le site **VentureScanner** qui était actualisé en juillet 2018[720]. Il organise le marché des startups de l'intelligence artificielle en 13 segments et évalue leur ancienneté et leur financement.

Et voici leur découpage, un peu restructuré pour le simplifier :

- **Plateformes** : machine learning, deep learning, réseaux de neurones, composants
    - **Plateformes de deep learning et machine learning** (passé de 256 startups en 2017 à 307 startups en 2018) : qui font avancer l'état de l'art côté algorithmie, avec des modèles prédictifs divers. Cela doit probablement intégrer des startups utilisant le machine learning dans des marchés verticaux.
- **Données** : prédictif, analytics, recommandation, en gros, ce qui ne concerne pas le cognitif, la vision et le langage
    - **Systèmes de recommandation** (passé de 102 à 112 startups) : pour prédire les comportements des utilisateurs. On en trouve notamment dans la musique, la vidéo et la restauration.
- **Vision** : recherche d'images, commande gestuelle
    - **Plateformes de vision** (passé de 191 à 218 startups) : technologies horizontales dans la vision artificielle comme les moteurs de recherche d'images ou les systèmes de tagging d'images.
    - **Applications vision** (passé de 214 à 258 startups) : appliquée à des marchés verticaux comme dans le commerce en ligne ou la santé.
    - **Reconnaissance de vidéos** (passé de 24 à 27 startups) : comme pour détecter les contenus protégés.
    - **Commande gestuelle** (passé de 60 à 64 startups) : à la frontière entre objets connectés et captation de mouvements et d'images.
- **Langage** : chatbots, traduction, extraction, recherche
    - **Traitement du langage** (passé de 304 à 346 startups) : solutions techniques dans le domaine pour comprendre le langage, le traduire, le résumer, le générer, etc.
    - **Reconnaissance de la parole** (passé de 163 à 181 startups) : avec des logiciels de reconnaissance de la parole, fournis en cloud ou en mode embarqué.
    - **Traduction vocale** (21 startups) : traduction « speech to speech » utilisable dans des contextes divers comme dans les chats vidéos.
    - **Assistants virtuels** (passé de 186 à 230 startups) : qui compend les chatbots qui sont déclinés en plateformes de chatbots et en chatbot pour marchés verticaux ou horizontaux, en b2b et b2c.

[720] Voir Artificial Intelligence Startup Highlights – Q2 2018 publié en août 2018. Ils suivaient 2262 startups dans l'IA dans 13 catégories sur plus de 70 pays représentant $38B de levées de fonds. On peut aussi télécharger gratuitement une base de 3465 startups sur le site d'Asgard, un autre organisme qui analyse le marché des startups, en s'appuyant d'ailleurs sur du traitement du langage automatisé pour extraire des informations structurée de textes libres récupérés sur Internet.



santé
sécurité
commerce
vente et marketing
fintechs
transports
HR

CB INSIGHTS

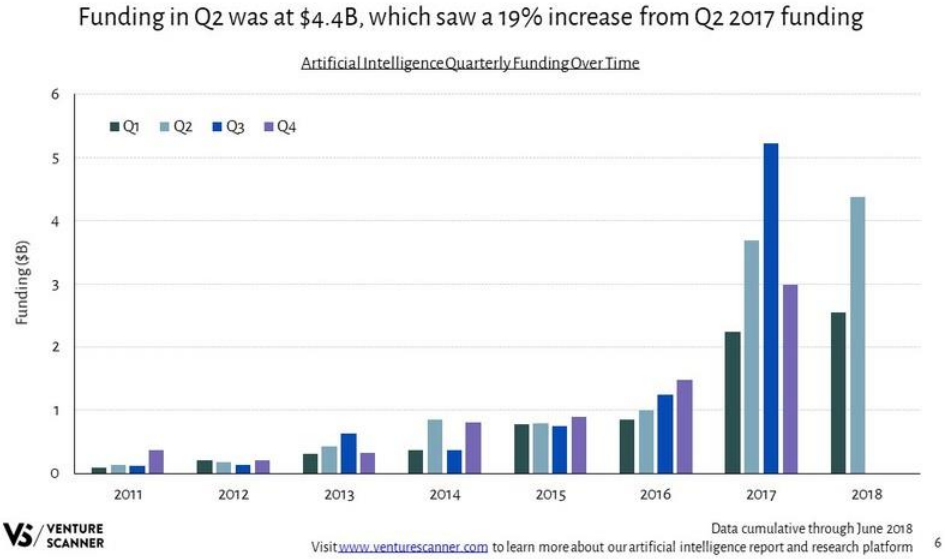
- **Robots** : plateformes, humanoïdes, services
    - **Robots** (passés de 168 à 189 startups) : du robot domestique au robot industriel.
- **Métier** : avec applications verticales (transports, santé, finance, ecommerce, juridique, agriculture) et horizontales (sécurité informatique, RH, marketing)
    - **Applications métier** (passé de 570 à 837 startups) : qui exploitent le machine learning et le deep learning en exploitant des données métier, comme la détection de fraude bancaire ou la génération de leads. La « heat map » de **CBInsights** de fin 2017[721] montre que les domaines d'action de ces startups sont en priorité dans la santé, la vente et le marketing puis les fintechs. L'IoT et la business intelligence sont aussi dans le top 5 mais ne sont pas spécifiquement verticaux.
    - **Applications contextuelles** (33 startups) : une catégorie un peu fourre-tout d'applications qui captent des données de l'environnement de l'utilisateur. Leur cartographie de juillet 2017 n'est pas bien à jour car Cleversense qui est dans cette catégorie a été acquis en 2011 par Google.

Dans la pratique, la frontière entre plateformes génériques et applications métiers est ténue. Nous avons par exemple vu dans le domaine de la santé des startups spécialisées dans l'imagerie médicale sur une seule pathologie (rétinopathie diabétique, dermatologie) ou d'autres qui en couvraient plusieurs.

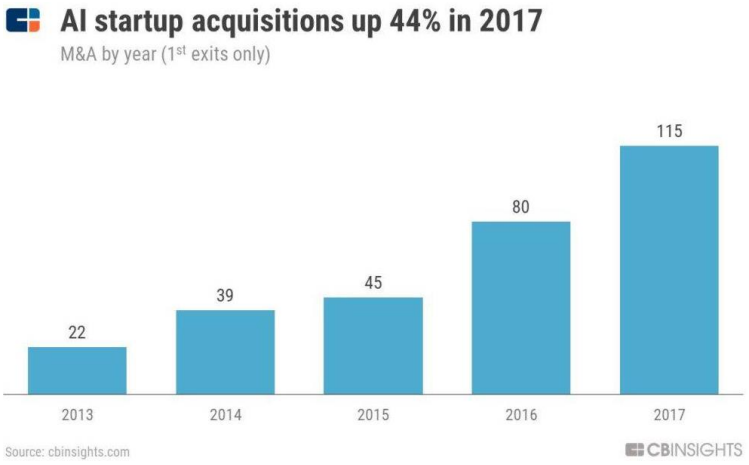
[721] Voir Top AI Trends To Watch In 2018 (38 slides).

Dans le traitement des données, des sociétés de plateformes sont en fait rapidement spécialisées dans le marketing ou la finance.

On constate une évolution à la hausse du financement des startups de ces secteurs. De 2009 à la mi-2018 comme dans le schéma *ci-contre* issu de **Venture Scanner** en août 2018.



L'ancienneté des startups de ce secteur est plutôt grande avec un bel étalement sur la date de création. Il y a certes un pic autour de 2012 pour celles qui ont levé des fonds. L'autre indicateur de l'activité autour des startups est celle des acquisitions de startups. La tendance était à la hausse jusqu'en 2017 et ne semble pas s'être calmée sur la première partie de 2018[722]. Et sur Q4 2017, quatre sorties étaient supérieures à $800M (*ci-dessous*).



Comme à chaque nouvelle vague technologique, celle de l'IA donne lieu à ce qui s'apparente à une bulle[723]. Une bulle en nombre de startups créées et en montants levés, surtout aux USA et en Chine. Plusieurs startups ont réussi à lever plus de $200M, mais ces montants sont encore modestes au regard de nombreuses unicorns ayant un patrimoine technologique assez réduit au départ comme les Pinterest de ce monde.

On peut qualifier ce phénomène de bulle pour au moins trois raisons spécifiques.

Tout d'abord, nombre de ces startups n'ont pas une « scalabilité » évidente, surtout lorsqu'elles sont b2b. C'est moins vrais pour les plateformes qui sont les vendeurs de pelles et de pioches du marché mais chez qui la concurrence sera rude et fermer émerger une toute petit poignée de leaders. Cette difficulté à « scaler » est amplifiée par l'IA car les données et les structures de données qui les alimentent sont souvent spécifiques à chaque client.

---

[722] Voir The Race For AI: Google, Intel, Apple In A Rush To Grab Artificial Intelligence Startups, février 2018.

[723] Voir Pour un chercheur d'Oxford, l'intelligence artificielle est une bulle spéculative, avril 2018, qui relate les propos de Michael Wooldridge, le dirigeant du département des sciences informatiques de l'Université d'Oxford. Il considère que beaucoup de fantasmes sont plaqués sur les capacités de l'IA. Elle progresse plus lentement qu'il n'y parait. Donc, les bénéfices de l'IA brandis par les startups sont survendus. Dans Why I don't invest in AI, juillet 2018, Peng Ong évoque le cas spécifique des startups d'IA s'attaquant à des marchés verticaux. La commoditisation de l'IA comme outil les rend rapidement moins différenciées. Les barrières à l'entrée sont similaires à l'habitude.

Seules les startups b2b exploitant des données indépendantes des entreprises clientes peuvent véritablement scaler. Cela pourrait être le cas de startups telles qu'Equals 3 et d'autres cas vus sur le marketing. L'IA est encore un véritable artisanat du logiciel. On est encore très loin d'une véritable industrialisation des procédés.

La seconde est que l'on est plus face à une tâche d'huile qu'à une bulle. A savoir que l'IA ne va bientôt plus être la spécificité de quelques startups avant-gardistes mais devenir courante dans presque toutes les startups et tous les éditeurs de logiciels. N'importe quel éditeur de solution logicielle – cloud ou pas – qui s'alimente de données va pouvoir les exploiter pour faire du machine learning, de la segmentation et des prévisions. La question se posera à chaque fois de la pertinence de le faire selon les volumes de données récoltés, mais on verra sans doute apparaître des méthodes permettant de se satisfaire de volumes de données limités. Bref, à terme, toutes les startups et tous les éditeurs de logiciels feront de l'IA et cela ne sera plus différenciateur à haut niveau !

S'y ajoute une autre tendance : nombre d'acquisitions de startups sont en fait des « *acquihires* [724] », notamment chez Google. Des chercheurs ont tout intérêt à créer une startup avant de se faire recruter par Google. Leur « sign-in » bonus sera ainsi plus élevé. Nombre d'acquisitions ont ainsi eu lieu pour des startups avec seulement un vague produit et pas vraiment de business model. C'est déjà arrivé par le passé mais le phénomène est plus marqué avec l'IA du fait de la rareté des compétences.

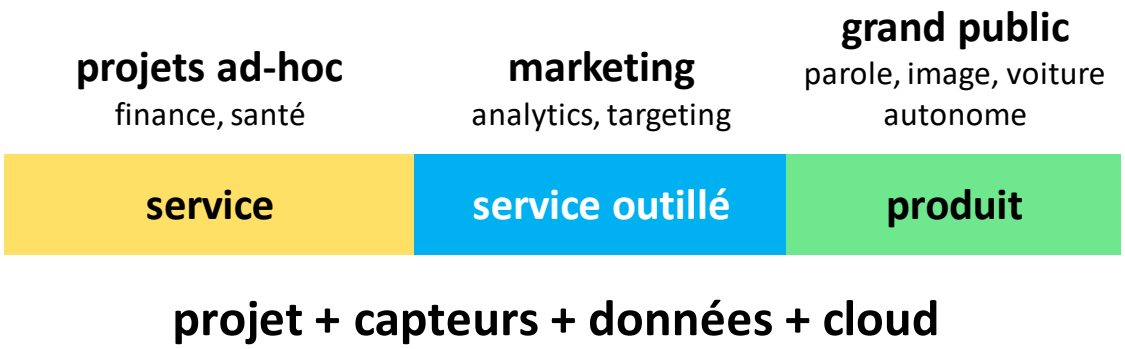
Evaluation

Quelles sont les caractéristiques d'une bonne startup faisant appel à de l'IA ? Ce sont quasiment les mêmes que pour les grandes entreprises évoquées dans la partie précédente. Ce petit inventaire est surtout destiné aux entreprises clientes mais est aussi utilisé par les investisseurs.

- **Talents** : une startup dans l'IA doit faire appel à des talents techniques variés avec des data-scientists, des développeurs maitrisant le paramétrage de systèmes de machine learning et deep learning, et les autres connaissances techniques classiques allant du back-end au front-end. Le design est aussi de la partie car les solutions logicielles se distinguent encore par là. La première chose à faire pour évaluer une startup consiste donc à consulter la biographie de ses fondateurs et de ses équipes techniques.

[724] Acquisitions based on hiring. Ce sont des acquisitions de startups qui sont faites pour recruter les développeurs et autres équipes techniques de la startup plus que pour récupérer un produit, leurs clients ou un marché.

- **Métier** : une startup connaît bien le métier de ses clients, ses contraintes, ses besoins et aspirations. Elle sait créer une solution qui s'intègre bien dans l'existant. Elle sait parler métier aux décideurs de l'entreprise.
- **Données** : il n'y a pas d'IA sans accès à des données pour entraîner ses modèles. Les données exploitées par la startup peuvent être de trois types : ouvertes et facilement disponibles sur Internet[725] (open data, ImageNet[726], WordNet[727], MNIST, TIMIT pour la reconnaissance de la parole en américain), collectées de manière exclusive à la startup via ses propres solutions matérielles et/ou logicielles, par exemple via ses objets connectés, ou provenant des systèmes d'information de ses propres clients. La différenciation de la solution provient généralement de la combinaison des trois sources. Une startup n'exploitant que des données ouvertes aura moins de barrières à l'entrée. Et pour accéder aux données des entreprises clientes, il faudra souvent faire du spécifique ce qui réduira les effets d'économie d'échelle de la startup. Autre question clé : où sont stockées les données ? Comment est géré le respect de la vie privée des utilisateurs pour les applications grand public ? La startup est-elle conforme à la RGPD en vigueur depuis le 25 mai 2018 ?
- **Produit** : est-ce que la solution est générique ou demande-t-elle d'adopter un mode projet lourd pour sa mise en œuvre chez chaque client ? Si on est en mode projet à chaque fois, on sera dans la catégorie des services outillés, la startup étant hybride entre startup produit et entreprise de services du numérique (ESN) avec peu d'économies d'échelle[728]. Et aussi, ne pas oublier d'avoir une démonstration du logiciel ! Dans l'IA, l'ergonomie est aussi importante que la fonction ! Et s'il n'y a pas de produit, c'est que l'on n'a pas affaire à une startup[729] !



[725] Voir la liste de jeux de données d'entraînement sur Wikipedia sachant qu'elles ne sont pas toutes publiées en open data : https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research#Speech.

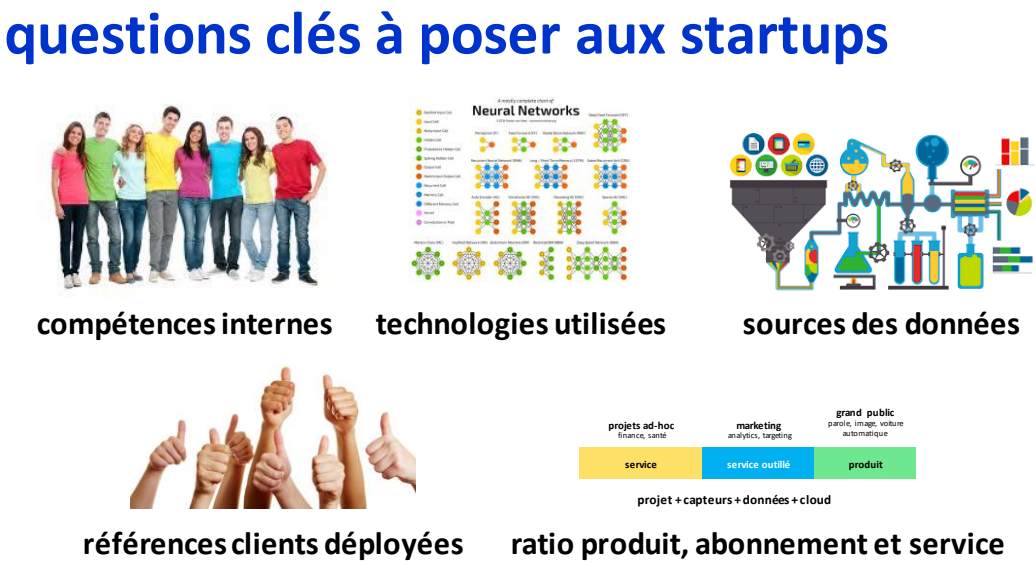
[726] La base ImageNet a été créée en 2009. Elle comprenait au départ 16 millions d'images associées manuellement à 100 000 mots de la base ouverte WordNet, ces mots étant organisés dans une arborescence.

[727] WordNet est une base de mots anglais et de synonymes.

[728] **Element.ai** (2016, Canada, $102M) est une startup de Montréal qui se focalise sur la création de produits pour différents marchés verticaux comme la finance et la supply chain, en faisant le pont entre la recherche et les entreprises clientes. Leur équipe de fondateurs comprend le fameux chercheur Yoshua Bengio du MILA. Leurs investisseurs comprennent Microsoft Ventures, Intel Capital et Nvidia. J'ai pu croiser son fondateur Jean-François Gamier à la résidence de l'ambassadeur du Canada à Paris le 10 octobre 2018. L'idée de la société est de créer des solutions génériques multi-clients architecturées autour de l'IA exploitant aussi bien des données sructurées et non structurées. Il reconnaissait que la création de produits était difficile en mode b2b et qu'il n'existait pas véritablement de startup « best in class » dans le domaine. L'un des manières de procéder est de trouver des idées communes à plusieurs clients que les clients sont prêts à adopter de manière non exclusive. La startup a déjà embauché 40 développeurs Français. Ils ont même des grands clients en France.

[729] De ce point de vue-là, la levée de fonds de 2M€ de Nabla.ai, créée par Alexandre Lebrun de Facebook est étonnante dans la mesure où la société se positionne comme une société de services. La levée s'explique par le pédigrée du fondateur qui avait déjà vendu une première startup à Facebook, wit.ai. Voir Nabla, la pépite IA créée par des anciens de Facebook et de MyLittleParis, 2018.

- **Technologies** : quand une startup indique avoir créé « son IA », il est bon de creuser un peu pour se faire expliquer le pourquoi du comment. Quels outils a-t-elle exploité pour créer sa solution ? Quelles méthodes de machine learning ou quels types de réseaux de neurones ? Est-ce que leur combinaison est originale, ce qui créera une différenciation, que l'on devra retrouver dans la performance de la solution ? Quelle est la partie algorithmique qui est spécifique à la startup ? Est-ce qu'elle a développé un savoir-faire spécifique dans l'assemblage de briques algorithmiques diverses ?
- **Business** : quel est le modèle économique de la startup ? Est-il récurrent ? Où sont les économies d'échelle ? Des questions assez classiques.
- **Financement** : c'est toujours le nerf de la guerre pour le développement de véritables startups, celles qui ont une grande ambition, notamment internationale. Nous avons vu dans les énumérations nombreuses de ce document que les startups US bénéficiaient souvent de financements importants, pouvant facilement dépasser les $30M, ce qui est plus rare en France et en Europe en général.



Ecosystème français de l'IA

Dans tous les nouveaux domaines technologiques, la France aspire collectivement à l'excellence et à la singularité. Pas celle de l'IA mais celle de la différenciation.

L'habitude est de mettre en valeur voire de monter en épingle l'excellence de nos chercheurs puis de nos ingénieurs, qu'ils soient restés en France ou expatriés dans de grandes entreprises internationales du numérique. Cela doit justifier notre place au soleil de la compétitivité scientifique, technologique et économique. Et nous avons Yann LeCun, le père des réseaux de neurones convolutionnels, qui œuvre chez Facebook, et Luc Julia, père de Siri, chez Samsung, etc !

Depuis 2016, diverses initiatives cherchent à valoriser les startups de ce nouveau marché et de mettre en avance l'excellence française et ses opportunités. Nous avons eu notamment le plan France IA du gouvernement en 2017, le rapport de la Mission Villani en 2018, la nomination d'un coordinateur de l'IA au sein de la DINSIC, Bertrand Pailhes en juillet 2018 et la création de l'association France Is AI d'ISAI en 2016.

Ce comportement est assez fréquent face à de nombreuses vagues technologiques : dans les jeux vidéos, dans les objets connectés, dans la cybersécurité, la robotique, et récemment, pour la Blockchain pour ne prendre que quelques exemples. Le timing de ces initiatives est variable. Il est en tout cas relativement tardif pour ce qui est de l'IA.

L'IA constitue bel et bien une opportunité de positionner le pays, à la fois ses entreprises numériques et les autres qui peuvent faire appel à l'IA pour innover et devenir plus compétitives. Mais nous sommes aveuglés par les mêmes erreurs de perspective que par le passé. Ce qui fera les forces et les faiblesses de ces startups n'est pas directement lié à l'IA mais plutôt générique. L'important est de savoir à quelle vitesse ces startups se financent, développent un prototype viable, génèrent leurs premières références clients et se déploient commercialement à l'étranger, surtout aux USA. Ces besoins sont génériques.

L'IA a cependant quelques caractéristiques qui ralentissent ces élans : les économies d'échelle sont bien plus grandes dans les activités tournées vers le grand public que vers les entreprises. Vers ces dernières, les startups de l'IA ont du mal à générer des économies d'échelle car chaque entreprise génère son propre projet avec ses spécificités et ses données propriétaires. Les économies d'échelle ne fonctionnent bien qu'avec des produits répondant à des besoins génériques, avec des plateformes de développement ou dans une certaine mesure avec des outils d'IA qui la démocratisent à une plus large audience, par exemple les TPE et PME.

Dans l'IA comme ailleurs dans le numérique, l'intégration est mondiale, pas française. Une startup ne peut devenir un leader mondial qu'en devenant une plateforme, c'est-à-dire, un produit extensible par des tiers. Les plateformes ne sont pas françaises pour le marché français, elles sont mondiales. L'avantage des startups américaines est dans la dimension de leur marché intérieur, qui conditionne à la fois leur surface commerciale initiale et leur capacité de financement. Pour les égaler, il faut trouver cette surface or la surface française est toujours comprise entre 1/7,5 (PIB) et 1/30 (financement dans le capital risque) vis-à-vis des USA. Il faut donc voir grand pour les meilleurs.

Cela correspond d'ailleurs à un progrès récent de l'écosystème entrepreneurial français. On compte chaque année une trentaine de startups qui dépassent les 10M€ de financement. Et elles s'orientent de plus en plus à l'international. Il faut continuer. L'investissement dans les startups atteindra 4Md€ en 2018, un niveau extraordinaire par rapport au 1Md€ qui sévissait il y a à peine quelques années. Mais ces investissements ont aussi grandi dans nombre d'autres pays. Dans l'ensemble de ce document, j'indique systématiquement les montants levés par les startups citées. Cela calme parfois un peu !

Le timing d'une mobilisation autour d'une nouvelle technologie définit indirectement le résultat que l'on peut en obtenir. Comme les principales plateformes du marché sont américaines, on devient des usagers de ces plateformes. Les économies d'échelle vont aux américains et le service et la personnalisation à nos acteurs.

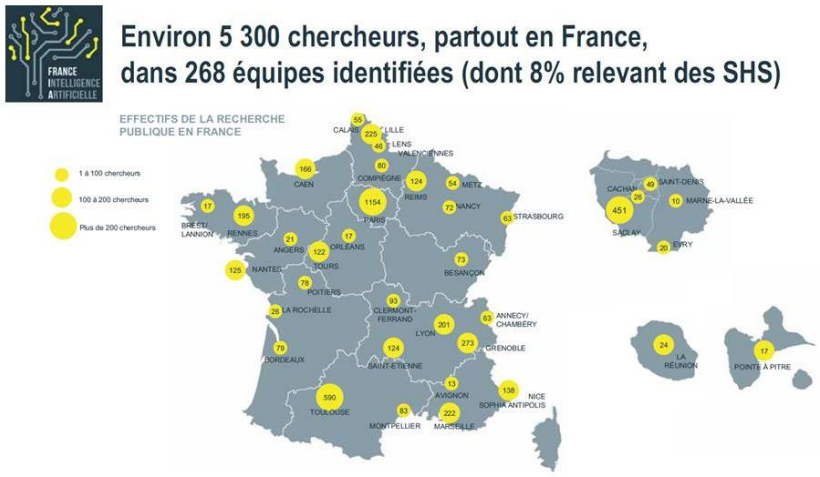
Qui plus est, l'IA va très rapidement devenir une commodité. Dans quelques années, il n'y aura plus de marché de l'IA. Il se confondra avec celui de l'informatique et du numérique, tant dans le grand public que dans les entreprises. La question clé sera alors son adoption par l'ensemble des entreprises et bien moins dans la création d'une industrie numérique de l'IA où les jeux auront été faits. On parlera alors de rattrapage des TPE et des PME, comme on a du le faire – et pas terminé – pour Internet.

Le développement de l'écosystème de l'IA, moins visible, concernera alors les prestataires de services et ceux qui proposent du service outillé. Nombre d'agences de communication, web agencies et autres entreprises de services numériques se mettent progressivement à l'IA et structurent leurs offres.

Recherche

Faisons d'abord un tour dans l'écosystème de la recherche en IA français. Il a été très bien inventorié dans le cadre du plan **France IA** du gouvernement, publié en mars 2017[730]. Il a probablement du évoluer depuis, mais à la marge.

Selon ce rapport, la recherche publique française serait disséminée dans plus de 220 équipes de recherche totalisant 5300 chercheurs, avec de nombreux projets collaboratifs associant laboratoires publics, universités et, parfois, entreprise privées.



	Country	Documents	Citable documents	Citations	Self-Citations	Citations per Document	H index
1	China	11894	11383	5204	3991	0.44	213
2	United States	8934	8427	3302	1582	0.37	437
3	India	4840	4548	906	433	0.19	108
4	Japan	2985	2847	616	295	0.21	145
5	United Kingdom	2461	2251	1214	439	0.49	228
6	Germany	2212	2018	759	380	0.34	186
7	France	1643	1482	542	216	0.33	162
8	Spain	1624	1406	562	234	0.35	145
9	Italy	1585	1387	752	398	0.47	143
10	South Korea	1564	1512	522	166	0.33	114

Les principaux organismes se focalisant sur l'IA sont l'INRIA, le CNRS et le CEA. La recherche en IA est soit « pure », soit appliquée à des domaines comme la santé.

Les chercheurs français seraient les plus prolixes en publications scientifiques, derrière les Américains et les Chinois qui dominent le secteur. La canadiens ne sont pas loin, aussi bien à Toronton qu'au Québec. En fait, la France serait 7e en publications scientifiques dans l'IA en 2017 [731], derrière UK, l'Allemagne, l'Inde, le Japon et bien évidemment, la Chine et les USA, talonnée de très près par l'Italie et l'Espagne, mais devant le Canada.

Pourtant, dans nombre d'inventaires de la presse anglo-saxonne, la France est royalement ignorée[732]. En fait, il existe plusieurs manières de classer les pays au niveau des publications scientifiques.

[730] Voir https://www.economie.gouv.fr/France-IA-intelligence-artificielle. Ce plan a été lancé par Axelle Lemaire puis repris par son successeur Christophe Sirrugue. Il faut considérer que le plan présenté en mars 2017 était un rapport d'étape. Après la présidentielle, le nouveau gouvernement a confié à Cédric Villani la mission de créer un plan complet, qui a été présenté fin mars 2018

[731] Source : https://www.scimagojr.com/countryrank.php?year=2017&category=1702. La France est également au septième rang sur la période 1996-2017. Cependant, en 2012 et 2013, l'Espagne était devant la France.

[732] Voir une petite revue de presse des pays leaders de l'IA qui ignore royalement la France : The AI race: 7 countries to watch out for, février 2018, 2017 in Review: 10 Leading AI Hubs, décembre 2017, 5 Countries Leading the Way in AI, janvier 2018, These Seven Countries Are In A Race To Rule The World With AI, décembre 2017 (qui cite l'Estonie et la Russie, mais pas la France !).

Selon une source de la même époque, mars 2017[733], la France compterait en fait 13 250 chercheurs en IA. C'est probablement une estimation un peu élevée. L'évaluation du nombre de ces chercheurs n'est pas évidente. Est-ce qu'un chercheur qui utilise de l'IA dans ses travaux dans la santé est un chercheur « en » IA ? Ce serait comme si on considérait qu'un chercheur en sciences des matériaux utilisant un ordinateur était un chercheur en logiciels.

Plus de 13 250 chercheurs en IA travaillent en France au sein du top 10 des laboratoires

Les laboratoires d'IA tricolores totalisant le plus de chercheurs au 10 mars 2017 :

Rang	Acronyme	Nom	Nombre de chercheurs
1	IRISA	Institut de recherche en informatique et systèmes	800
2	IRIT	Institut de recherche en informatique de Toulouse	651
3	STICC	LAB-STICC	560
4	LIP6	Laboratoire d'informatique de Paris 6	517
5	LIG	Laboratoire d'informatique de Grenoble	500
6	LORIA	Laboratoire lorrain de recherche en informatique et ses applications	500
7	LS2N	Laboratoire des sciences du numérique de Nantes	450
8	LIRMM	Laboratoire d'informatique, de robotique et de microélectronique de Montpellier	444
9	CRISTAL	Centre de recherche en informatique, signal et automatique de Lille	430
10	LIST	Institut List - CEA	360

La recherche privée en IA en France serait dominée par les laboratoires ouverts par des entreprises étrangères : Sony, Facebook, Huawei, Microsoft, Rakuten et Xerox. Sachant qu'ils ne sont pas dédiés à 100% à de la recherche en IA. Il faudrait y ajouter le laboratoire d'IA installé à Grenoble par le Coréen Naver pendant l'été 2017[734].

Une R&D IA privée française active accompagnée par 9 acteurs de renom

Sur les neuf entreprises qui ont ouvert, à ce jour, un laboratoire d'intelligence artificielle dans l'Hexagone, **trois sont françaises : Criteo, Michelin et Orange**.

Origine des entreprises ayant implanté leur R&D IA en France :

Acronyme	Nom	Entreprise rattachée	Pays d'orgine
Criteo Labs	Criteo Labs	Criteo	France
CSL Sony	Laboratoire scientifique Sony computer	Sony	Japon
Factolab	Factolab	Michelin	France
FAIR Paris	Facebook AI research Paris	Facebook	Etats-Unis
Huawei MASL	Mathematical and algorithmic sciences lab	Huawei Technologies	Chine
MSFT	Centre Microsoft recherche-Inria	Microsoft corporation	Etats-Unis
Orange Labs	Orange Labs	Orange	France
RIT Paris	Rakuten institute of technology Paris	Rakuten	Japon
XRCE	Centre de recherche Europe de Xerox	Xerox	Etats-Unis

Dans le privé, Criteo, Orange et Michelin sortent du lot. Orange a toujours eu un gros laboratoire de recherche. Actuellement dirigé par Nicolas Demassieux et situé principalement dans les locaux Orange Gardens à Chatillon sous Bagneux, avec quelques autres équipes à Rennes et Lannion, il comprend diverses équipes qui exploitent des briques d'IA dans leurs recherches. Cela porte notamment sur des aspects très opérationnels du déploiement à venir des réseaux 5G.

En juin 2018, Criteo annonçait investir 20M€ sur trois ans pour créer son laboratoire de recherche en IA à Paris, qui complète ses équipes produits qui comprennent déjà 300 ingénieurs et développeurs. Comme il se doit, ce Criteo AI Lab sera très focalisé sur le machine learning. Le laboratoire est dirigé par une chercheuse indienne, Suju Rajan, que l'on voit fréquemment intervenir dans les conférences en France.

L'INRIA a publié courant 2016 un excellent livre blanc qui décrit ses priorités et projets dans l'IA[735]. De manière assez classique, les projets portent sur le langage, la vision et la robotique. L'INRIA planche aussi beaucoup sur l'IA symbolique avec le web sémantique, les neurosciences et sciences cognitives ainsi que sur la programmation par contrainte. Elle s'intéresse à la protection de la vie privée ainsi qu'aux applications de l'IA dans la santé.

---

[733] Voir Intelligence artificielle en France : la carte des laboratoires de Lélia De Matharel du JDN, mars 2017. Ces 13250 chercheurs ne sont pas mentionnés dans la page recherche du site France Is AI. Selon le Ministère de l'Enseignement Supérieur et de la Recherche, il y aurait environ 428 600 équivalents temps plein chercheurs en France en 2017.

[734] Ce laboratoire occupe 70 chercheurs à comparer aux 200 chercheurs de Naver situés en Corée du Sud. Naver a en fait repris clé en main l'ancien centre de recherche de Xerox de Grenoble créé en 1993. Le laboratoire est spécialisé en vision artificielle, mobilité et traitement du langage. Naver a aussi lancé un incubateur à Station F où il héberge notamment Videolabs, et a financé à hauteur de 200M€ le fonds d'investissement Korelya Capital lancé par Fleur Pellerin.

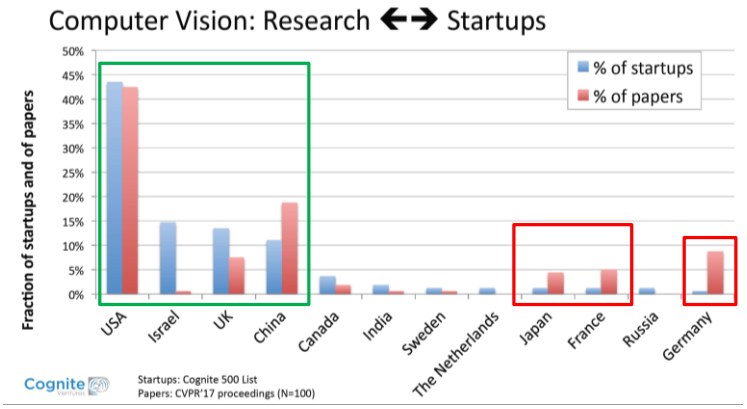
[735] Voir Intelligence artificielle, dédis actuels et l'action d'Inria, 2016 (82 pages).

Il est difficile de caractériser les spécificités de la recherche française en IA. Elle est multi-domaine sans spécialisation apparente. On peut cependant y distinguer une force dans l'IA symbolique et la logique formelle, dans le traitement du langage ainsi, en filigrane, qu'un souci de créer des solutions d'IA responsables et éthiques[736].

Le défi pour ces chercheurs et leurs autorités de tutelle est de trouver des applications marchés de leurs travaux. En consultant la liste des participations d'**IT-Translation**[737] qui est l'un principaux financeurs de projets issus de l'INRIA, on constate que l'IA est souvent en filigrane de ces projets.

Les travaux des chercheurs en IA n'aboutissent pas naturellement à des projets entrepreneuriaux. Ne serait-ce parce qu'il faut une couche de traduction entre ces réalisations et leurs applications et que les innovations des startups résultent souvent de la combinaison de plusieurs méthodes et techniques[738].

Cognite Ventures faisait un point en 2017 de la création de startups dans un domaine précis, la vision artificielle, au regard des publications scientifiques dans le domaine[739]. Il fait ressortir deux types de pays : ceux qui valorisent bien leurs travaux de recherche dans des startups (USA, UK, Chine) et ceux qui au contraire publient bien plus qu'ils ne créent de startups (France, Japon, Allemagne). Israël est un cas avec très peu de publications et beaucoup de startups.



On peut difficilement extrapoler cela mais force est de constater que les startups issues de la recherche ne sont pas très nombreuses en France dans l'IA.

En France, la recherche dans l'IA semble mieux financée côté civil, même s'il est difficile de le vérifier par les chiffres. On ne s'en plaindra pas. A ceci près que la R&D militaire US a une qualité : elle est orientée vers des objectifs pratiques selon des cahiers des charges. De son côté, la recherche civile française fonctionne plutôt de manière très décentralisée.

Startups

France Is AI a inventorié 305 startups françaises opérant dans l'IA dont une bonne vingtaine a des implantations à l'étranger et notamment aux USA[740]. En élagant la liste et en supprimant les startups acquises et les doublons, on tombe à un total de 298 startups[741].

---

[736] Cela se retrouve notamment dans les travaux de Laurence Devillers, du CNRS-LIMSI qui portent sur le langage, sur la détection des émotions et sur l'éthique de l'IA.

[737] Voir le portefeuille de participations d'IT Translation.

[738] Pour apprécier la difficulté à transformer la recherche en IA en solutions métiers, vous pouvez par exemple consulter les actes de la dernière conférence ICML sur le machine learning qui s'est tenue début août 2017 en Australie : https://2017.icml.cc/Conferences/2017/Schedule.

[739] Voir Does vision research drive deep learning startups?, août 2017. Le chart de cette page ne correspond pas exactement à celui de l'article. Je n'arrive pas à retrouver sa source.

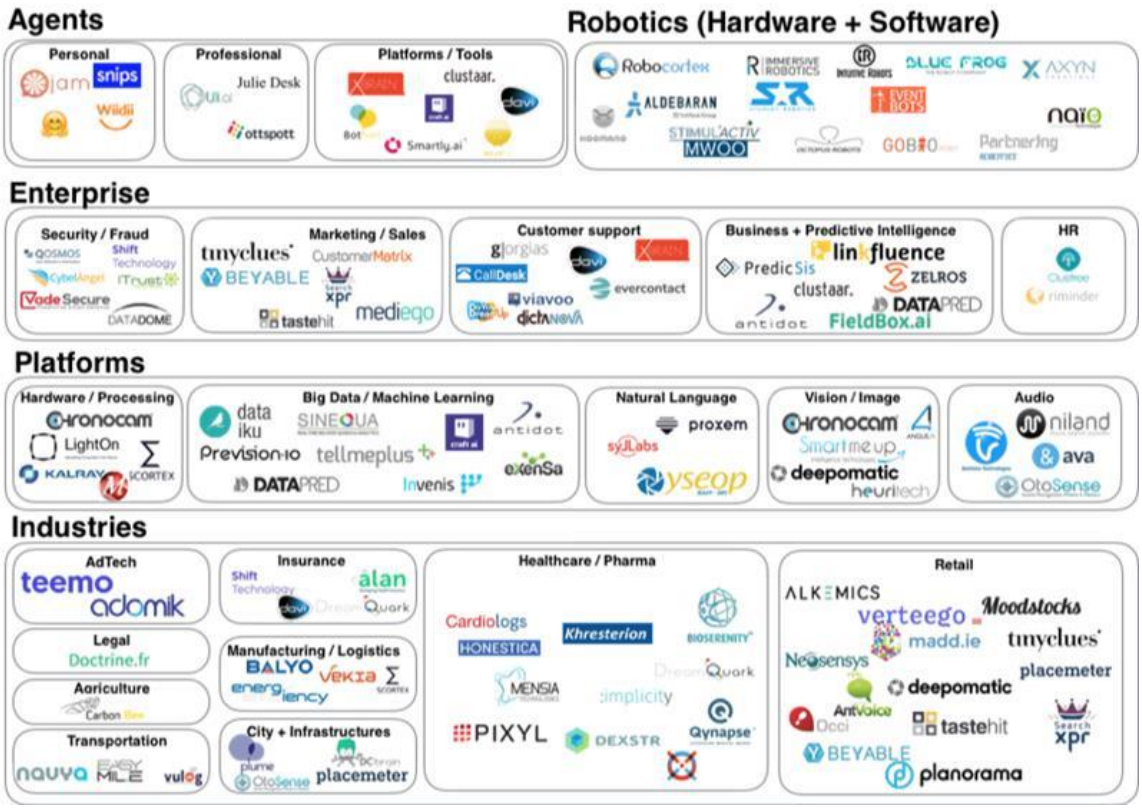
[740] Source du chart : https://franceisai.com/startups, octobre 2018.

[741] La liste est déclarative. Les startups s'inscrivent d'elles mêmes. Avec deux écueils : des startups qui n'utilisent pas forcément de briques logicielle d'IA ou celles qui sont des prestataires de services en mode projet et n'ont pas de produit, et tout au plus une boîte à outils plus ou moins propriétaire.

La France se retrouve en seconde position européenne selon les critères habituels de nombre de startups et de levées de fonds. Cette position a l'air d'être la même, ni plus ni moins, que dans les startups en général[742].

La diversité de leur activité est très grande avec deux marchés verticaux qui émergent, celui de la santé et celui du retail. Elles sont à dominantes b2b. On y trouve un grand nombre de chatbots avec une difficulté à distinguer les sociétés de services des véritables plateformes technologiques de création de chatbots[743].

En octobre 2018, les 18 startups de ce panorama ayant levé plus de $10M étaient Algolia ($74M), Dataiku ($45,7M), Shift Technology ($39,8M), Navya ($38M), Vulog ($31M), Sophia Genetics ($29,75M), Alkemics ($27,6M), Tinyclues ($25,4M), Snips ($21M), Databerries ($17,8M), Vekia ($17M), Easymile ($15M), Delair ($14,5M), Arterys ($14,2M), Saagie ($12,51M), Clustree ($11,55M), PriceMatch ($10,38) et ReCommerce ($10M). Même si ce n'est pas le seul indicateur de la croissance, c'en est qui est intéressant à suivre pour détecter les « scale-ups », ces startups qui arrivent à bien grandir au-delà de leur phase d'amorçage.



Entreprises

Nous en avons cité quelques unes dans la grande partie de cet ebook qui est dédiée aux marchés verticaux : nombre de grandes entreprises ont adopté l'IA dans leur informatique interne. C'est plus visible dans les entreprises qui gèrent de gros volumes de données comme dans les télécoms, la finance, l'assurance, la distribution et la santé. Les utilities ont aussi fort à faire avec l'IA pour optimiser et maintenir leurs infrastructures.

---

[742] Donc, de quoi relativiser ce que l'on peut lire dans Pourquoi la France est un poids lourd de l'intelligence artificielle, de Géraldine Russell. octobre 2018.

[743] Voici un petit inventaire de startups de chatbots françaises : AskHub, Askr.ai, Botfuel, Botmind, Clustaar, Golem, Haapie, Heek, Hellomybot, Hubware, Opla, Owi, Posos, Smartly, Snapdata, Synapse Développement, VisualBot.ai, Voxist, Wiidii et xBrain. Ca fait du monde !

Dans les transports, la plupart des acteurs, des constructeurs aux opérateurs, peuvent aussi faire appel à l'IA. Il en va de même de l'ensemble du complexe militaro-industriel. Il est cependant peu évident de jauger de la vitesse à laquelle ces entreprises adoptent l'IA. Je n'ai pas encore trouvé de benchmark international. En gros, les équipes innovation et les entreprises les plus hardies sont déjà à plein pied dans l'IA. La majorité cependant en sont encore au stade de la réflexion et des projets pilotes. La courbe d'apprentissage est naturellement lente.

Associations

La France est le pays des associations loi 1901 et cela commence à se refléter dans les associations autour de l'IA.

A commencer par l'**Association Française pour l'Intelligence Artificielle** (AFIA), créée en 1993, et qui fait la promotion de la recherche en IA. Elle regroupe donc surtout des chercheurs. Comme il y a toujours du pour et du contre, nous avons aussi une **Association contre l'IA**, l'AFCIA, créée en 2015 et qui vise simplement à interdire à l'échelle mondiale toute recherche sur l'IA[744]. C'est l'équivalent anti-OGM ou anti-nucléaire appliqué à l'IA.

On retrouve une dichotomie avec deux associations : d'un côté **FranceIsAI** et de l'autre, le **Hub France IA**, deux mouvements qui ont des visées différentes. **FranceIsAI** est une émanation du fonds d'investissement ISAI. C'est désormais une branche de l'association France Digitale qui associe entrepreneurs et investisseurs dans les startups françaises. Elle produit un inventaire des startups en France régulièrement mis à jour (cf. page précédente) et organise notamment l'excellente conférence FranceIsAI chaque année à l'automne, qui met en valeur à la fois les chercheurs et les entrepreneurs du secteur.

Le **Hub France IA** a été créée par des chercheurs et industriels de l'IA français. Il a pour vocation de fournir à l'écosystème une plateforme de ressources visant a stimuler l'accélération de l'adoption de l'IA par les différents acteurs de l'économie, surtout les entreprises, notamment via des groupes de travail thématiques, associant grands groupes, startups et laboratoires de recherche en IA, une offre d'initiation et de formation à l'IA, des rencontres et du networking entre membres et des aides à la constitution de dossiers de financement publics pour des programmes de R&D.

Il existe aussi une association l'**IA pour l'Ecole**, qui vise à créer des ponts entre les pionniers de l'IA et tous les acteurs de l'éducation.

Un point clé : l'IA ne doit pas être qu'une affaire d'hommes. Comme dans le développement logiciel, on y trouve malheureusement plutôt une toute minorité de femmes alors que ces technologies vont conditionner le futur de l'humanité et du travail. Pourtant, on trouve plein de femmes remarquables dans l'IA, comme le montre cet inventaire US[745]. D'où l'intérêt d'initiatives telles que **Women in AI**, une association mondiale avec une branche en France qui fédère les femmes travaillant dans le secteur de l'intelligence artificielle et qui cherchent à attirer d'autres femmes dans le domaine. Des initiatives communautaires ont été également lancées en région comme **Lyon in AI**, piloté par Amélie Cordier, CTO de la startup de logiciels en robotique **Hoomano**.

Partenariats internationaux

En juin 2018, Emmanuel Macron et Justin Trudeau lançaient un projet franco-canadien en IA, en l'espèce, un groupe international d'étude réunissant des experts indépendants (scientifiques, gouvernements, industrie et société civile). Il s'agit visiblement surtout d'un groupe de réflexion sur la régulation de l'IA, pas vraiment d'un laboratoire de recherche. Le principal livrable ? La création d'un groupe de travail qui devra définir les contours de ce projet. Bref, on a annoncé un objet dont on ne connaissait pas la nature au moment de son lancement.

---

[744] Voir On peut être contre l'intelligence artificielle par principe de Irénée Régnauld, publié sur Uzbek&Rica en janvier 2017

[745] Voir Meet these incredible women advancing AI research, Topbots, mai 2017.

IA et société

Nous allons sortir ici des considérations techniques et d'entreprises pour aborder la place de l'IA dans la société et tenter de traiter quelques questions clés.

Quels boulevesements annonce-t-elle, notamment dans le travail et l'emploi ? Quelles craintes et espoirs soulève-t-elle ? Comment la politique et l'Etat s'en emparent-ils ? Comment la réglementation pourrait-elle évoluer ? Quel est le rôle des entreprises de ces points de vue-là ? Pourquoi parle-t-on d'éthique de l'IA ? Que penser des prophètes de la singularité ? C'est l'objet de cette partie.

Craintes sur l'IA

L'IA génère-t-elle plus de craintes que les machines à tisser, les chemins de fer, l'aviation commerciale, l'énergie nucléaire ou les OGM au moment de leur apparition ? Il est difficile de comparer des époques différentes mais l'IA est en tout cas entrée dans le club plutôt fermé des technologies qui font peur.

Une bonne part de ces peurs provient de la science fiction plus que de la science, ainsi que des effets d'annonce enjolivés à la moindre avancée du deep learning et à une conception étriquée de la notion même d'intelligence humaine. On use et on abuse trop facilement de la loi de Moore, simplifiant à volonté la notion même d'intelligence humaine pour la comparer à celle des machines. On confond métiers et tâches, calculs et décisions, automatisation et outillage. Et le moutonisme médiatique fait son effet, comme la titraille ci-dessous l'illustre.

Et cela se retrouve dans les enquêtes d'opinion. Selon l'enquête iLife de l'agence de communication BETC du groupe Publicis réalisée en février 2017 auprès de 12 169 personnes de plus de 18 ans, les français sont plutôt pessimistes sur l'impact de l'IA sur la société, sur l'emploi et sur la vie en général[746].

L'IA conserve un côté magique qui permet de faire prendre des vessies pour des lanternes à de vastes audiences, y compris celles qui sont éduquées. Mais, même en étant prudent et conservateur, on peut estimer que l'IA aura un impact aussi important que les 35 années de vagues d'innovations numériques qui viennent de se succéder.

---

[746] Voir l'étude iLife de BETC (slides). L'IA est bonne pour la société : Chine : 89%, USA : 53%, France : 33%. L'IA va créer des emplois : Chine : 59%, USA : 48%, France : 34%. L'IA impactera positivement nos vies : Chine : 45%, USA : 54%, France : 64%. L'IA nous libérera : Chine : 78%, USA : 46%, France : 36%.

C'est au minimum une grande vague de « logiciels 2.0 » qui est lancée à vive allure. Mais au même titre que la situation d'aujourd'hui était difficile à prédire il y a 35 ans, celle des 35 ans l'est tout autant. La raison est simple : les progrès de l'IA sont non seulement incertains d'un point de vue scientifique et technologique, mais s'y ajoutent les habituelles dimensions économiques et sociales qui jouent toujours le rôle de filtre entre l'univers du possible et ce qui devient disponible.

Une bonne part des craintes provient aussi de la propension à projeter sur les robots et l'IA nos propres défauts. La vision antropomorphique de l'IA est à l'origine d'une bonne partie de nos fantasmes et peurs sur l'IA.

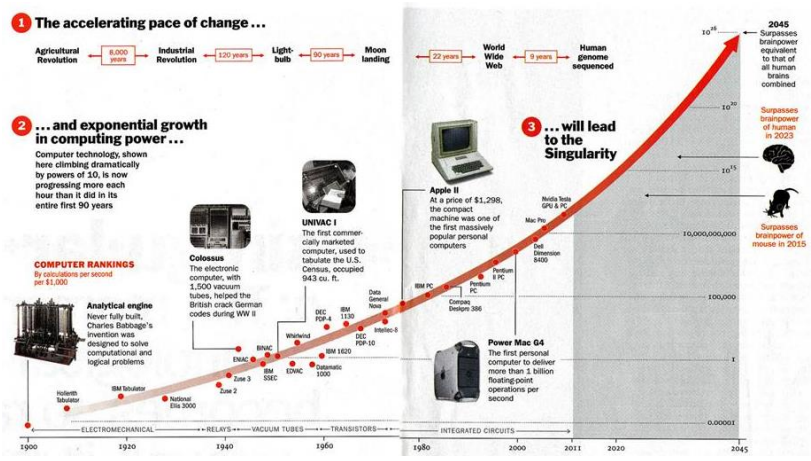
Elle est justifiée dans la mesure où une bonne partie du savoir exploité par l'IA est d'origine humaine. C'est en limitant cet antropomorphisme à la fois dans nos projections et dans la création de systèmes à base d'IA que l'on peut revenir sur un terrain de confiance vis-à-vis de cette dernière.

Risques

Avec l'IA, l'Homme a peur d'être dépassé par ses propres créations, peur de perdre le contrôle de son devenir, à la fois intellectuel et dans la maîtrise du monde physique. Pourtant, les évolutions technologiques passées n'avaient pas cet impact. L'Homme est d'ailleurs déjà largement dépassé par ses propres créations depuis longtemps[747], d'abord du côté de la force physique, puis de calcul, de mémoire et enfin, de traitement. Les machines mécaniques dépassent la puissance humaine depuis des lustres. On s'y est habitué et on s'en accomode. Ce sont des outils. Il n'y a que dans le raisonnement abstrait que les machines ont encore du mal à se développer.

Cette peur est alimentée par la perspective de voir émerger d'ici à peine quelques décennies une IA généraliste quelque peu mythique (AGI = artificial general intelligence) omnisciente, omnipotente et contrôlant tout notre monde physique, et qui pourrait en retour nous asservir. Cette peur s'appuie sur une extrapolation des capacités actuelles de l'IA.

Elle s'appuie notamment sur une vision très simpliste d'une l'application ad-vitam de la loi de Moore (*ci-contre*). L'intelligence humaine ne se réduit pas à un nombre de transistors ou de neurones. Plus on rentre dans le détail du fonctionnement des neurones, plus on découvre leur complexité. Depuis les prédictions de Ray Kurzweil en 2005 sur la singularité, la complexité apparente des neurones a été multipliée par 100[748].



Plus les futurologues de l'IA ont de véritables connaissances scientifiques en IA et en neurosciences, moins ils en ont peur. Nombre d'entre eux considèrent même que l'AGI, tout comme la singularité, sont des mythes.

---

[747] Et pas besoin d'en ajouter avec des annonces comme New AI Can Write and Rewrite Its Own Code to Increase Its Intelligence, février 2017,

[748] C'est un ordre de grandeur aussi approximatif que ses prévisions !

Vous avez certainement entendu parler des craintes de l'IA manifestées par Bill Gates, Elon Musk, Stephen Hawking, Yuval Harari ou Henri Kissinger[749]. Ils sont certes connus et respectés[750]. Mais ils ont ceci de commun de ne pas être des spécialistes de l'IA[751]. Yann LeCun et Luc Julia sont eux beaucoup plus optimistes, ou moins pessimistes, comme vous le voulez[752]!

Dans des scénarios de prospective dignes des meilleures dystopies de science fiction, la première AGI génèrerait d'elle-même une ASI (Artificial Super Intelligence) qui prendrait le contrôle de la planète et annihilerait toutes les autres AGI, physiquement, via un contrôle direct des infrastructures, ou via divers « hacks ». Cette anticipation est une vue de l'esprit très centralisatrice[753].

Certes, Google domine l'Internet occidental avec son moteur de recherche et Facebook domine les résaux sociaux occidentaux mais Internet reste assez distribué en l'état.

Comme avec l'usage de nombreuses autres technologies (automobile, armes à feu, …), le vrai danger est l'Homme lui-même. C'est lui qui décide ce qu'il va en faire[754].

C'est lui qui créé les règles utilisées par l'IA. Elles peuvent respecter les lois de la robotique d'Asimov… ou pas ! L'IA est plus ou moins dangereuse selon ce que l'Homme lui donne à faire et à décider[755]. Si un chatbot s'entraîne en discutant avec des internautes racistes, il deviendra lui-même raciste comme l'a malheureusement expérimenté Microsoft avec Tay en 2016. Bref, l'enfer, c'est l'Homme, pas l'IA !

L'IA présente des risques bien plus prosaïques, comme toutes les technologies numériques : dans sa sécurité et dans ce qu'on lui demande de faire ! La sécurité d'un système d'IA peut être compromise à plusieurs niveaux : dans les réseaux et le cloud, dans les capteurs, dans l'alimentation en énergie.

---

[749] Voir AI could mean the end of human history, de Henry Kissinger, dans The Atlantic, juin 2018.

[750] Exemple avec cette conférence « Où va nous mener d'intelligence artificielle » de Georges Mitaut, en décembre 2016, ingénieur en télécommunications retraité, qui ressasse les lieux communs habituels sur l'IA et la singularité sous forme d'une revue de presse bien slidisée. Au milieu de sa présentation et à propos de Google se trouve un logo de TensorFlow. Le conférencier indique alors que cela fait de l'IA mais qu'il n'a aucune idée de la manière dont cela fonctionne ! Ce n'est pas une exception ! La majorité des conférenciers sur l'IA font de la prospective alors qu'ils ne connaissent même pas les mécanismes actuels de l'IA et l'état de la recherche fondamentale. C'est consternant !

[751] Mais parfois, ils connaissent l'IA et ne sont pas d'accord. C'est le cas de Sergei Brin. Voir Google Cofounder Sergey Brin warns of AI's dark side tandis qu'Eric Schmidt, l'ancien chairman de Google, trouve qu'Elon Musk est trop pessimiste. Voir Google billionaire Eric Schmidt: Elon Musk is 'exactly wrong' about A.I. because he 'doesn't understand', mai 2018 ainsi que Elon Musk is Wrong, AI won't kill us all de Toby Walsh, septembre 2017.

[752] Voir l'interview Pour moi, l'IA n'existe pas de Luc Julia dans CBNews en juin 2018. Et pour Yann LeCun : IA : « Les scénarios à la Terminator n'existent pas », estime Yann LeCun dans Numerama, octobre 2017, Yann LeCun, de Facebook : « l'intelligence artificielle va sauver des vies », dans Le Monde, septembre 2017 puis Yann LeCun : "L'intelligence artificielle a moins de sens commun qu'un rat", dans Sciences et Avenir, janvier 2018.

[753] Voir cette intéressante analyse des biais cognitifs de ces prévisions alarmistes : The Seven Deadly Sins of Predicting the Future of AI de Rodney Brooks, septembre 2017.

[754] Voir Will AI kill us all after taking our jobs? de Fabio Ciucci publié en juillet 2017 ainsi que AI (Deep Learning) explained simply publié en juin 2017.

[755] Dans « Life 3.0 », Max Tegmark évoque la thèse de Nick Bostrom dans « Superintelligence » selon laquelle le but ultime d'un système est indépendant de son intelligence, qui fournit les moyens d'atteindre un objectif complexe indépendamment de sa nature. L'intelligence humaine est ainsi déjà mise au service du bien commun et de la science ou du mal sous toutes les formes. Aux USA, ce dernier cas relève des « smart assholes » ! A la fin de son ouvrage, Max Tegmark indique que pour garder le contrôl de l'IA, il faut qu'elle apprenne, adopte puis conserve les objectifs qui lui ont été assignés par les Humains.. Voir Life 3.0 by Max Tegmark review – we are ignoring the AI apocalypse, septembre 2017, une review du livre de max Tegmark par Yuval Harari.

Les bases de connaissances peuvent aussi être induites en erreur par l'injection d'informations erronées dans des images qui visent à altérer son comportement, par exemple dans le cadre d'un diagnostic médical complexe. On peut imaginer l'apparition d'anti-virus spécialisés pour les logiciels de deep learning.

Comme le code et les logiciels régissent de plus en plus notre vie, ce sont les règles qu'ils exécutent qui organisent la société. L'IA est aussi basée sur des règles (pour les moteurs de règles et les réseaux d'agents) et sur le mimétisme des sens et comportements humains (pour le deep learning).

Les dangers potentiels d'une hypothétique AGI sont surtout liés aux interactions entre les machines l'exécutant et le monde extérieur. Un robot logiciel n'est pas dangereux s'il tourne dans une machine isolée. Il peut le devenir s'il contrôle une arme de destruction massive ou des infrastructures stratégiques dans le monde extérieur et qu'il est programmé par des forces maléfiques ou simplement inconscientes.

La capacité à débrancher une AGI est devenue un thème de recherche des plus sérieux. Google annonçait en janvier 2016 qu'il travaillait sur la notion de « kill switch » de l'IA sans que l'on en connaisse d'ailleurs la nature. On sait juste que ce sont des chercheurs de DeepMind qui étudient des scénarios d'interaction entre robots et hommes dans des situations sous contrainte multiples : assurer une tâche d'un côté et réagir à des imprévus d'autre part.

Le "kill switch" de l'AGI qui permettrait de la déconnecter si elle devenait dangereuse devrait surtout porter sur sa relation avec le monde physique. Même si les films de science fiction tels que Transcendance rappellent que rien n'est sûr de ce côté là et que la tendance à tout automatiser peut fournir un trop grand contrôle du monde réel aux machines. Reste à trouver le « kill switch » pour une IA fortement distribuée, comme Skynet !

Je m'étonne toujours de notre capacité à construire des paquebots, porte-containers et porte-avions de 300 à 400 m de long et pesant plus de 100 000 tonnes. Un tableur compte plus vite que n'importe quel champion de calcul mental, ce depuis 1979. Autant la capacité de traitement parallèle d'un cerveau humain est impressionnante, autant sa capacité de stockage est limitée.

Une simple clé USB de 32 Go peut contenir plus de textes que ce que nous lisons, écrivons, entendons et disons pendant toute notre vie [756]!

Et ce que nous retenons ne fait qu'environ 1 Go ! Plus les outils numériques stockent l'information et sont faciles à utiliser pour l'interroger, moins on la retient.

L'IA est aussi anxyogène car elle peut générer des systèmes pérennes dans le temps. Ses processus d'apprentissage bénéficient de la mémoire presque infinie des machines. L'IA serait donc immortelle, tant que ses systèmes de stockage ne défaillent pas. On peut se rassurer en rappelant qu'un disque dur peut planter à tout bout de champ au bout de cinq ans et qu'un disque SSD actuel ne supporte au mieux que 3000 cycles d'écriture ! Mais leur remplacement robotisé est tout à fait possible dans des datacenters. Enfin, les data centers ont besoin d'énergie et ils sont encore rares à être autonomes de ce point de vue-là[757].

---

[756] Je me suis amusé à faire le calcul suivant : une personne qui vit 85 ans représentant 31 025 jours, pendant lesquels elle va lire 100 pages par jour, en écrire 20 par jour, et parler ou écouter parler pendant 8 heures par jour à raison de 200 mots à la minute va générer ou être exposée à 41 Go de données textuelles. C'est évidemment un cas extrême. Pour les gens moins bavards, moins lecteurs et moins producteurs, cela va tomber largement à moins de 10 Go. Ce qui ne fait pas grand-chose ! Qui plus est, on ne retient qu'une toute petite portion de tout cela. Donc, notre mémoire verbale ne fait probablement qu'à peine qu'environ 1 Go, avec cependant de nombreuses ramifications. Cette évaluation à la louche n'intègre pas la mémoire visuelle et auditive, qui est bien plus dense. La mémoire est notamment limitée par la durée de notre vie et la vitesse d'accumulation de connaissances et expériences.

[757] Un data center alimenté par sa propre centrale nucléaire serait très dangereux. Et il n'existe pas de data centers alimentés entièrement par des panneaux solaires photoltaïques. Leurs onduleurs permettent en général de tenir quelques heures ou journées sans alimentation électrique.

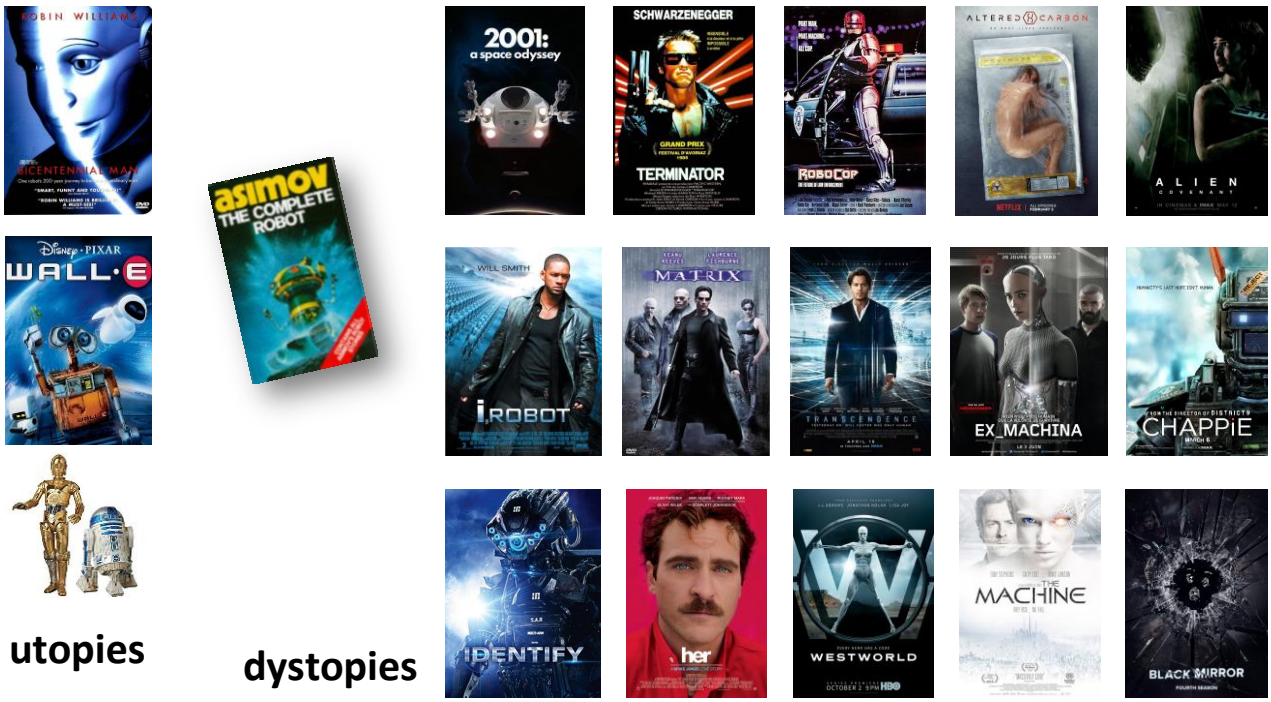
Mais les forces obscures humaines veillent au grain. Quelle sera l'arme de destruction massive à base d'IA ?

L'autre peur, plus court terme, et que nous étudierons plus loin concerne les évolutions des métiers qui, soit disparaitront, soit deviendront bien plus productifs grâce à l'IA. C'est une crainte économique et sociale plus qu'une crainte de perte de contrôle de l'IA par l'Homme.

Pessimistes

La première source de pessimisme concernant l'impact de l'IA sur l'humanité est la science fiction. Si l'on observe la production cinématographique des dernières décennies, les dystopies prennent largement le dessus des utopies. L'utopique Bicentenial Man en 1999 a été un flop tandis que tous les Terminator et son Skynet ont été des blockbusters[758].

Ceci étant dit, les films racontant la vie heureuse de familles de cadres aisés avec trois enfants sont moins fréquents que les films d'horreur ou les policiers en tout genre. Les morts parmis les agents de la CIA dans certains films et certaines séries d'espionnage (Jason Bourne, 24, Scandal, etc) permettraient de remplir plusieurs murs de mémorial de la CIA ! Et quand la science se fait fiction, elle déraille presque systématiquement et l'Homme en perd le contrôle, comme dans la récente série Westworld[759].



[758] Même si on passe de la dystopie à l'utopie pour ce qui est du rôle de ce robot à partir du second film. Avec un bon robot contre un mauvais robot, la dualité bien/mal humaine reproduite dans les machines !

[759] Voir Westworld, as reviewed by scientists, roboticists, researchers par Jennifer Bisset, juillet 2018, et aussi The incredible, unbelievable, rapidly advancing future, and the end of The End de Ben Edwards, octobre 2017, définit quelques-uns des scénarios catastrophes d'une IA dont l'Homme perdrait le contrôle.

Malgré tout, les grands auteurs de science fiction tels qu'Isaac Asimov, ont été prescients dans un grand nombre de leurs fictions. Ils l'ont d'ailleurs été plus sur l'impact sociétal des technologies de la communication que sur les technologies dans d'autres domaines (transports, santé, …).

Les alertes sur les risques de l'IA gagnent en écho lorsqu'elles proviennent de personnalités scientifiques et entrepreneuriales. L'astrophysicien **Stephen Hawking** n'hésitait pas à prophétiser en 2014 que lorsque l'IA dépassera l'intelligence humaine, ce sera la dernière invention humaine, celle-ci ayant ensuite pris entièrement le pas sur l'espèce humaine [760]!

Il reprenait à son compte une citation d'Irwin John Good de 1965[761] selon laquelle la machine ultra-intelligente sera la dernière invention que l'homme aura besoin de créer (*ci-dessous*). Pour autant, si Hawking s'y connaît bien en trous noirs, il n'est pas forcément spécialisé en réseaux de neurones et deep learning.

9. Conclusions

These "conclusions" are primarily the opinions of the writer, as they must be in a paper on ultraintelligent machines written at the present time. *In the writer's opinion then*:

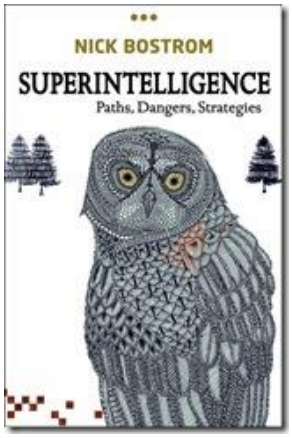
It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make, since it will lead to an "intelligence explosion." This will transform society in an unimaginable way. The first ultraintelligent machine will need to be ultraparallel, and is likely *to be achieved with the help of a very large artificial neural net.*

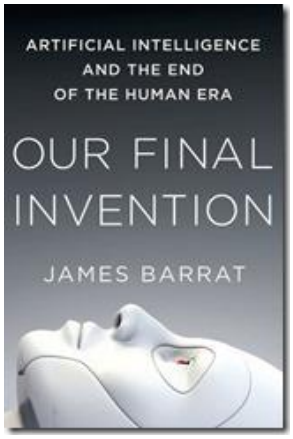
Cette thèse se retrouve décrite dans le menu dans de nombreux ouvrages, comme **Superintelligence** de Nick Bostrom paru en 2014 ou dans **Our Final Invention, Artificial Intelligence and the End of the Human Era** de James Barrat, paru en 2015.

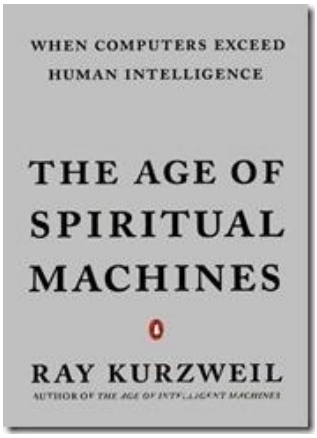
Ces prédictions partent du principe que l'on arrivera un jour à créer une machine superintelligente dont la puissance croitra de manière exponentielle et qui contrôlera toutes nos destinées du fait de l'hyperconnexion des infrastructures physiques et des objets de la vie courante.

Le cofondateur de Sun Microsystems, **Bill Joy**, avait été l'un des premiers à alerter l'opinion en 2000[762], tirant la sonnette d'alarme sur les dangers des progrès technologiques dans l'IA, les nanotechnologies et les biotechnologies.

C'était bien avant la fin du premier séquençage complet du génome humain qui avait coûté une fortune[763]. Bill Joy était en fait effrayé des perspectives avancées par Ray Kurzweil qu'il avait rencontré dans une conférence en 1998 et après avoir lu son **The age of spiritual machines**, paru six ans avant **The singularity is near**.







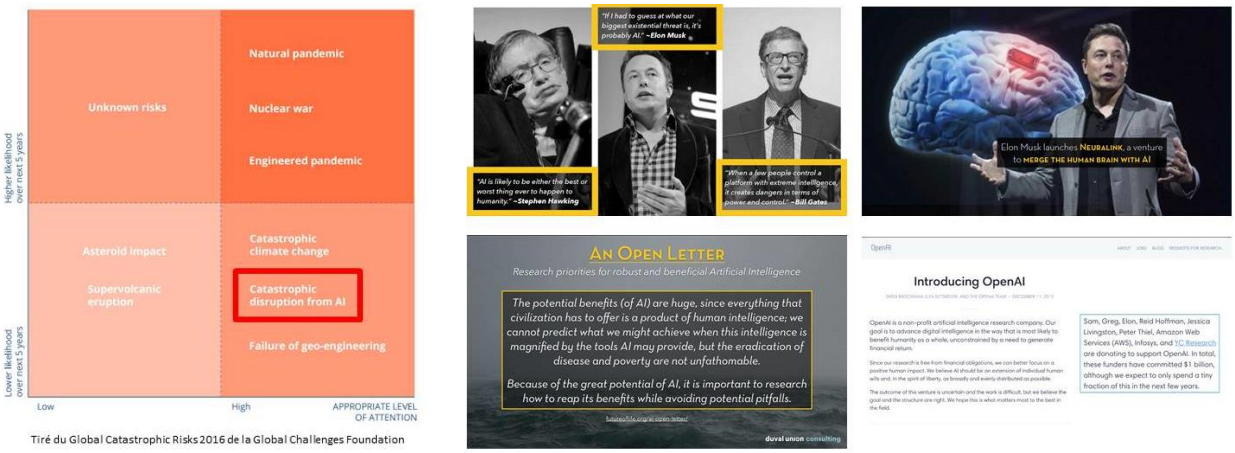
[760] Voir Stephen Hawking warns artificial intelligence could end mankind, mars 2014.

[761] Publiée dans Speculations Concerning the First Ultraintelligent Machine, 1965 (30 pages).

[762] Dans Why the future doesn't need us, Bill Joy, Wired, 2000.

[763] On y apprend d'ailleurs qu'il avait rencontré Jacques Attali et que ce dernier avait indirectement influé le cours des événements de Java !

S'en est suivie une grosse décennie de calme côté alertes. Après Stephen Hawking en 2014, Bill Gates et Elon Musk ont repris le flambeau de Bill Joy en 2015 pour demander une pause technologique et une réflexion sur les limites à ne pas dépasser avec l'intelligence artificielle comme avec la robotique. Pause de quoi précisément ? Ce n'était pas bien clair. Peut-être pour les rares startups bien financées qui planchent sur l'AGI comme Numenta. Mais aucune pause n'a eu lieu. Par contre, cela a lancé le débat sur l'explicabilité de l'IA et poussé les gouvernements et leaders d'opinions à s'intéresser à l'éthique de l'IA.



Tiré du Global Catastrophic Risks 2016 de la Global Challenges Foundation

Il existe même des instituts de recherche qui planchent sur la question des risques de l'IA, tels le **Center for the Study of Existential Risk** de Cambridge et le **Future of Humanity Institute** d'Oxford.

Les dangers perceptibles de l'IA sont à l'origine de la création d'**OpenAI**, fin 2015, une initiative visant non pas à créer une IA open source – les grands logiciels de l'IA sont déjà presque tous open source - mais à surveiller et analyser ses évolutions. Il s'agit d'une ONG créée par Elon Musk qui veut s'assurer de manière asez manichéenne que l'IA fasse le bien et pas le mal. C'est une vision assez naïve du fonctionnement du capitalisme. OpenAI milite en tout cas pour une régulation de l'IA[764].

Si par exemple, les méthodes de recrutement se mettent à utiliser de plus en plus l'analyse automatique de personnalité via les vidéos enregistrées par les candidats, on est déjà aux limites de l'éthique, mais ce n'est pas de l'AGI qui menace l'humanité. OpenAI est doté de $1B de financement et fait de la recherche et de la formation. Dans certains cas, leurs actions semblent un peu éloignées de l'objectif initial, comme dans ce benchmark de jeux vidéo opposant des joueurs humains et artificiels.

D'autres initiatives semblables ont vu le jour comme **Partnership on AI**[765], une association créé en 2016 et rassemblant comme membres fondateurs tous les GAFAMI (Google, Facebook, Amazon, Microsoft, IBM, et Apple qui les a rejoints en janvier 2017).

[764] Voir OpenAI Director Shivon Zilis explains why AI requires oversight now, novembre 2017.

[765] Voir https://www.partnershiponai.org/.

L'association est présidée par Eric Horvitz, le patron de Microsoft Research et Mustafa Suleyman de Google, le co-fondateur de DeepMind.

Elle débat des questions soulevées par l'IA et des meilleures pratiques à adopter pour en mitiger les risques. Quand on y regarde de près, cette association prend la forme d'une organisation de lobbying avec les méthodes associées : des thématiques à défendre, l'appel à des experts divers, l'organisation de débats et un pied dans la porte des politiques pour éviter des dérives réglementaires gênant l'innovation. Eric Horvitz promeut de son côté l'utilisation de l'IA pour le bien de l'humanité[766].

AINow est une autre initiative lancée mi 2016. Elle associe Kate Crawford (Microsoft Research) et Meredith Walker (Google Open Research Group). Basée à New York, elle est est focalisée sur l'impact de l'IA sur les droits civiques, sur l'emploi, les biais et la sécurité des infrastructures. Elle a publié un premier rapport en 2016[767].

Sous couvert de bonne gestion du principe de précaution, ces initiatives des GAFAMI sont à analyser sous la loupe des pratiques habituelles du lobbying[768]. Elles visent aussi à calmer les peurs et à assurer les pouvoirs publics qu'une autorégulation de l'IA est possible par les acteurs de l'industrie.

Cela vise surtout à éviter que ces derniers s'immiscent dans la stratégie de ces grands acteurs. Et dans le cas où il viendrait à l'idée des pouvoirs publics de réguler l'IA d'une manière ou d'une autre, d'être prêt avec des propositions compatibles avec leurs stratégies. C'est de bonne guerre mais il ne faut pas être dupe !

Enfin, Elon Musk a aussi lancé fin 2016 sa startup **Neuralink** dont l'objectif est de relier l'IA à l'homme pour éviter d'en perdre le contrôle[769], via des nano-électrodes directement implantées dans le cerveau et capables d'activer sélectivement les neurones. Ce projet est loin de pouvoir tenir ce genre de promesses.

Il servira probablement surtout à améliorer l'état de l'art du traitement de certaines pathologies neurodégénératives diverses, qui ne nécessitent pas d'agir au niveau de neurones individuels[770]. Et quand bien même le système fonctionnerait, il rendrait l'homme vulnérable au hacking d'une IA piratée par d'autres hommes mal intentionnés ! C'est donc une solution tout à fait bancale.

L'autre personnalité inquiète du futur de l'IA est le chercheur anglais **Stuart Russell**. Précisément, il veut réguler l'usage d'armes robots. Il a fait produire la vidéo de fiction glaçante Slaughterbots sur un hypothétique futur où de drones armés d'explosifs ciblent des dissidents dans un pays occidental.

---

[766] Voir son support de présentation, bien documenté d'études de cas d'usages positifs de l'IA : AI in Support of People and Society, juin 2016 (81 slides).

[767] Voir The AI Now Report - The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term, A summary of the AI Now public symposium, hosted by the White House and New York University's Information Law Institute, juillet 2016 (37 pages) et Why AI is still waiting for its ethics transplant, de Scott Rosenberg dans Wired, novembre 2017.

[768] Exemple récent : le lancement du challenge AI Impact doté de $25M pour récompenser des projets qui font de l'IA for good. Voir Google pledges $25 million toward AI solutions for social issues, octobre 2018.

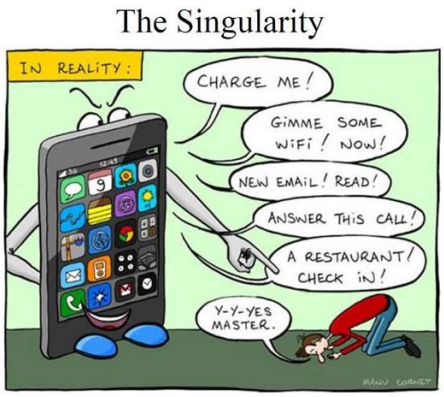
[769] L'idée est inspirée des neural laces de l'auteur de science fiction Iain M. Banks. Voir The novelist who inspired Elson Musk If you want to understand where society is heading, read the novels of Iain M. Banks, Silicon Valley's favourite author, de Tim Cross, mars 2017.

[770] On peut aussi imaginer des solutions visant à activer les neurones de l'hippocampe qui est une sorte de gatekeeper de la mémoire. C'est lui qui transfère la mémoire court terme au sein du cerveau limbique vers la mémoire long terme du cortex en périphérie du cerveau.

Elle était diffusée à l'occasion d'une conférence de l'ONU sur le sujet. Stuart Russell veut interdire ce genre de drones et toute forme de robot tueur[771]. Jusqu'au jour où un régime dictatorial décidera d'en créer et d'en utiliser ! Le pire est que la technologie présentée dans la vidéo ne relève pas de la science fiction et est disponible dès aujourd'hui.

Dans le top du top des prévisions, il y a aussi celle-ci qui prévoit qu'eu 2042 une IA va créer son dieu et sa propre bible. Elle vient d'un certain **Anthony Levandowski** qui a déposé les statuts d'une organisation religieuse "The Way of the Future » qui ambitionne de créer un dieu basé sur une IA qui ferait le bien de la société[772]. 2042, ça doit être une blague de Douglas Adams ! Si cela se trouve, le risque de l'IA sera plus basique. L'abus d'IA pourrait nous enfermer dans notre passé et limiter la sérendipité de notre vie qui deviendrait toute lisse.

On continuerait à se fier de plus en plus aux machines pour organiser notre vie comme on le fait déjà avec les outils de communications qui nous interrompent sans cesse. Cet abandon progressif de notre libre arbitre serait consenti volontairement et progressivement. L'IA abuserait de nos faiblesses pour nous orienter dans nos choix comme le fait déjà la publicité qui exploite sans vergogne nos pulsions[773].



la vie des utilisateurs est déjà largement régie par les logiciels et les plateformes

Optimistes

Les optimistes semblent moins nombreux. On y trouve bien évidemment les singularistes dont le pape actuel, **Ray Kurzweil**, anticipe l'émergence d'une AGI autour de 2030-2040 en nous promettant monts et merveilles qui peuvent nous encourager à procrastiner sur la résolution des problèmes d'aujourd'hui (réchauffement climatique, surpopulation, inégalités, …).

La procroastination a d'ailleurs lieu sans qu'il en soit responsable. Donald Trump n'a certainement pas lu Kurzweil[774] !

Pour le sceptique éclairé **Piero Scaruffi**, nous sommes tiraillés entre deux extrêmes de science-fiction avec des pessimistes qui pensent que l'IA va tous nous tuer et des singularistes qui estiment qu'elle va nous rendre immortels.

---

[771] Voir UN to host first talks on use of 'killer robots', novembre 2017.

[772] Voir An AI god will emerge by 2042 and write its own bible, will you worshipt it?, octobre 2017.

[773] Voir Les péchés capitaux dans le marketing et la vente, Olivier Ezratty, 2011, Comment l'intelligence artificielle va rendre notre vie terriblement ennuyeuse, de Céline Deluzarche, juin 2018 ainsi que Evoque Re-Engineering Humanity de Brett Frischmann et Evan Selinger, avril 2018.

[774] Ray Kurzweil est présenté selon les circonstances comme le directeur de la recherche de Google, son patron de la R&D en IA, directeur de l'engineering quand ce n'est pas « chief futurist ». Alors que les deux principales équipes d'IA de Google, Google Brain chez Google X et celle de DeepMind ne dépendent pas de Kurzweil. Il n'a rien produit ou annoncé depuis son arrivée chez Google en 2012. On sait qu'il planche sur le traitement du langage avec une équipe d'une vingtaine de personnes, à comparer aux 300 personnes qui travaillent chez DeepMind. Il travaillerait à la création d'un chatbot qui répondrait autoamtiquement à nos emails à notre place. Voir What is Ray Kurzweil up to at Google ? Writing your emails, février 2017. Il a fait une apparition au CEBIT 2017 où il a passé le plus clair de son temps à rappeler les effets de la loi de Moore et de ses déclinaisons dans d'autres domaines. Et d'évoquer quelques avancées dans la compréhension du fonctionnement du cerveau. On a surtout pu remarquer qu'il a maintenant une chevelure de quadra alors qu'il était quasiment chauve avant. Grâce aux plus de cent pillules qu'il prend chaque jour depuis des années pour prolonger sa durée de vie, à une greffe ou à une perruque ? Ray Kurzweil a aussi piloté le projet Talk to Books, permettant de poser des questions « à un livre » qui va y trouver les réponses … ou pas.

Au milieu de l'échiquier des optimistes se situent des personnalités telles que **Mark Zuckerberg** qui estime que l'Homme sera raisonnable dans ses usages de l'IA[775] et puis **Ginni Rometti** d'IBM qui recommande de ne pas avoir peur des robots[776]. D'autres comme **Sarah Kessler** prévoient que la transformation des métiers générera un nouvel équilibre, pas forcément moins bon que l'actuel, et qu'il n'y a pas lieu de s'inquiéter[777].

Les optimistes sont aussi souvent les véritables spécialistes de l'IA qui voient de près l'ingratitude de la discipline et estiment en général que l'on est très éloigné de l'AGI et de l'ASI. La plupart des auteurs qui prédisent une ASI ne sont en effet pas des spécialistes de l'IA[778] ! Les entrepreneurs de l'IA sont d'ailleurs fort agacés des prévisions délirantes concernant l'IA[779].

Le principal moyen de s'en éloigner est de faire la distinction entre l'association de l'intelligence humaine avec sa chair et ses sens, et n'importe quelle forme d'intelligence intégrée dans une machine dénouée de cette chair et de ces sens.

L'un des écueils principaux des prévisions pessimistes est leur anthropomorphisme[780], tourné dans le mauvais sens [781]! Le deep learning exploite souvent de l'expertise d'origine humaine, dans DeepMind AlphaGo aussi bien que dans les systèmes de reconnaissance d'image en imagerie médicale, Il en va de même pour IBM Watson en cancérologie qui exploite la littérature scientifique d'origine humaine sur le sujet. L'IA applique une force brute sur une vaste base de données d'intelligence humaine.

L'apprentissage supervisé fonctionne par imitation et utilise des tags d'origine humaine. Et l'apprentissage non supervisé, comme dans les premières phases des réseaux de neurones convolutionnels exploite en bout de course de l'apprentissage supervisé. Idem pour les modèles génératifs qui appliquent des styles d'origine humaine à divers contenus.

La soi-disante créativité des réseaux de neurones génératifs s'appuie toujours sur la créativité d'origine humaine qu'elle ne fait que répliquer machinalement et de manière prédictible ! Bref, en matière d'intelligence, l'IA imite le plus souvent celle de l'homme qu'elle met en forme et peut gérer en masse qu'elle n'en génère ex-nihilo.

Pour s'écarter de cet antrophomorphisme, on peut adopter la posture de **Kevin Kelly**, auteur du best seller « The inevitable », et qui considère que l'IA doit être considérée comme « alien »[782]. On peut aussi écouter les envolées lyriques du philosophe **Grady Booch** qui explique pourquoi il ne faut pas avoir peur de l'IA (vidéo), que l'Homme entraînera à ne pas lui nuire.

---

[775] Voir Intelligence artificielle : Zuckerberg (Facebook) n'a pas peur, dans ZDNet, février 2016.

[776] Voir At Davos, IBM CEO Ginni Rometty Downplays Fears of a Robot Takeover de Claire Zillman dans Fortune, janvier 2017.

[777] Voir The optimist's guide to the robot apocalypse de Sarah Kessler, mars 2017.

[778] La situation s'inverse entre Elon Musk (pessimiste) et Mark Zuckerberg (optimiste) lorsque le premier accuse le seccond d'ignorance sur l'IA. Et on revient au point initial lorsque le roboticien Rodney Brooks contredit Elon Musk. Dans This famous roboticist doesn't think Elon Musk understands AI, TechCrunch, juillet 2017.

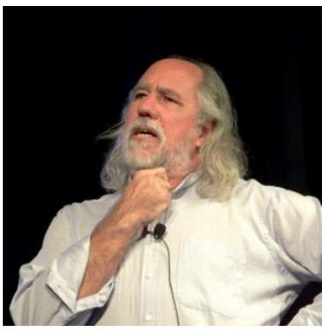
[779] Voir par exemple La supériorité de l'intelligence artificielle : l'arnaque du siècle, de Denis Fagès, fondateur de VideoTelling, juillet 2018. Et aussi We're told to fear robots. But why do we think they'll turn on us? The robot uprising is a myth de Steven Pinker, février 2018 ainsi que Beware the AI delusion de Gary Smith, octobre 2018.

[780] Comme cette petite tribune de Stéphane Mallard, Digital Evangelist, comme quoi l'IA sera capable de tout faire et d'être créative. L'IA, plus créative que l'Homme ?, dans l'ADN, avril 2017. Et on se fait avoir régulièrement par l'IA ! Voir 'Artificial Intelligence' Has Become Meaningless de Ian Bogost dans The Atlantic, mars 2017.

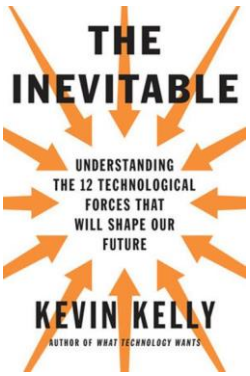
[781] Encore une savoureuse citation de Piero Scaruffi : "*In private conversations about "machine intelligence" i like to quip that it is not intelligent to talk about intelligent machines: whatever they do is not what we do, and, therefore, is neither "intelligent" nor "stupid" (attributes invented to define human behavior). Talking about the intelligence of a machine is like talking about the leaves of a person: trees have leaves, people don't. "Intelligence" and "stupidity" are not properties of machines: they are properties of humans. Machines don't think, they do something else. Machine intelligence is as much an oxymoron as human furniture. Machines have a life of their own, but that "life" is not human life.*".

[782] Voir Le mythe de l'IA surhumaine de Rémi Sussan, mai 2017. Kevin Kelly décrit cela lui-même dans Wired en avril 2017 : The myth of a superhuman AI.

conférence sur les usages "for good" de l'IA



Grady Booch



De son côté, le chercheur français **Jean-Gabriel Ganascia**, auteur du « Mythe de la singularité » (2016) dénonce avec justesse la construction de mythes autour de l'IA et de la singularité[783].

On peut aussi s'amuser de la crédulité de ceux qui ont avalé la création de **Rocket AI** (2016), une startup développant un réseau de neurones rappelant ceux de Numenta et baptisé « Temporal Recurrent Optimal Learning » (TROL). Il s'agissait d'une grosse blague de potaches de l'IA[784] soulignant la crédulité de l'écosystème de l'innovation.

Autre méthode, se rassurer une fois encore avec les écrits de **Piero Scaruffi**[785]. Ce dernier cherche à démontrer que la singularité n'est pas pour demain. Il s'appuie pour cela sur une vision historique critique des évolutions de l'intelligence artificielle. Il pense que les progrès de l'IA proviennent surtout de l'augmentation de la puissance des machines, et bien peu des algorithmes (ce qui serait à nuancer…). Il relativise les performances actuelles de l'IA, montées en épingle par les entreprises, les experts et les médias.

Selon lui, l'Homme a toujours cherché une source d'intelligence supérieure, qu'il s'agisse de Dieux multiples ou unique, de Saints ou d'extra-terrestres. La singularité et les fantasmes autour de l'IA seraient l'incarnation d'une nouvelle forme de croyance voire même de religion, une thèse aussi partagée par Jaron Lanier, un auteur anticonformiste qui publiait **Singularity is a religion just for digital geeks** en 2010[786].

Piero Scaruffi prend aussi la singularité à l'envers en avançant que l'ordinateur pourra fort bien dépasser l'Homme côté intelligence parce que les technologies rendent Homo Sapiens plus bête[787], en le déchargeant de plus en plus de fonctions intellectuelles, la mémoire en premier et le raisonnement en second !

Selon lui, le fait que les médias numériques entraînent les jeunes à lire de moins en moins de textes longs réduirait leur capacité à raisonner. A tel point qu'il devient impossible d'expliquer les effets de la baisse d'attention du fait de cette dernière[788] !

---

[783] Voir Technologie : peut-on se défaire des promesses et des mythes ?, une excellente revue de lecture de l'ouvrage de Jean-Gabriel Ganascia ainsi que de l'ouvrage collectif « Pourquoi tant de promesses » dirigé par Marc Audétat, par Hubert Guillaud, juin 2017.

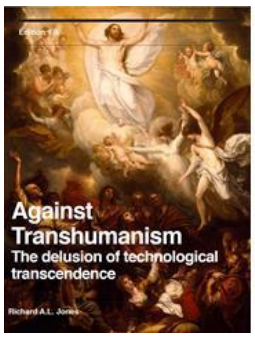
[784] Voir Rocket AI: 2016's Most Notorious AI Launch and the Problem with AI Hype, décembre 2016. Le site de Rocket AI n'est d'ailleurs pas moins documenté que celui de nombreuses startups de l'IA.

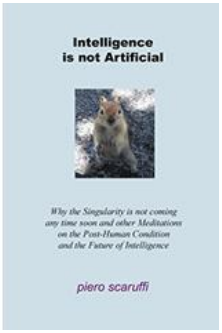
[785] Comme Demystifying Machine Intelligence.

[786] Voir Singularity Is a Religion Just for Digital Geeks, 2011.

[787] Thèse partagée par **Daniel C. Dennett**, pour qui le véritable danger n'est pas dans les machines plus intelligentes que l'homme mais plutôt dans le laisser-aller de ce dernier qui abandonne son libre arbitre et confie trop de compétences et d'autorité à des machines qui ne lui sont pas supérieures.

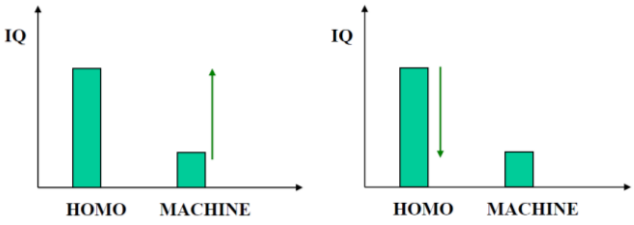
[788] "*I am worried that people's attention span is becoming so short that it will soon be impossible to explain the consequences of a short attention span. I don't see an acceleration in what machines can do, but i do see a deceleration in human attention… if not in human intelligence in general*", dans "Intelligence is not artificial".

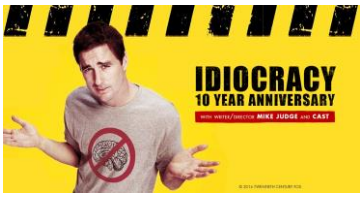




On peut d'ailleurs le constater dans les débats politiques qui évitent la pensée complexe et privilégient les simplismes à outrance. J'aime bien cet adage selon lequel l'intelligence artificielle se définit comme étant le contraire de la bêtise naturelle. Cette dernière est souvent confondante et rend le défi de la création d'une intelligence artificielle pas si insurmontable que cela dans un bon nombre de domaines.

on peut vérifier le test de Turing en rendant les machines plus intelligentes où les gens plus bêtes…





"***Know-nothingism*** *is the insistence that there are simple, brute-force, instant-gratification answers to every problem, and that there's something effeminate and weak about anyone who suggests otherwise*"
Paul Krugman, New York Times, Aug 2008

Pour Piero Scaruffi, en tout cas, l'intelligence artificielle est d'ailleurs une mauvaise expression. Il préfère évoquer la notion d'**intelligence non humaine**. Une bonne approche qui souligne la complémentarité de l'IA et des Hommes.

Il pense aussi qu'une autre forme d'intelligence artificielle pourrait émerger : celle d'hommes dont on aura modifié l'ADN pour rendre leur cerveau plus efficace. C'est un projet du monde réel, poursuivi en Chine où sont séquencés des milliers d'ADN humains pour identifier les gènes de l'intelligence ! Histoire de réaliser une (toute petite) partie des fantasmes délirants du film Lucy de Luc Besson !

L'intelligence humaine cumule la capacité à créer des théories expliquant le fonctionnement physique du monde et à mener des expériences permettant ensuite de les vérifier. Parfois, cette vérification s'étale sur un demi-siècle à un siècle, comme pour les ondes gravitationnelles ou l'existence du boson de Higgs. Cette capacité de théorisation et d'expérimentation de long terme n'est pour l'instant pas accessible à une machine, quelle qu'elle soit. Les machines ne se posent pas encore de questions existentielles sur leur relation au monde qui les entoure.

Emplois et IA

Prenons maintenant un peu de recul sur la robotisation en marche des métiers liée aux avancées de l'intelligence artificielle vues jusqu'à présent. Cette robotisation n'a pas besoin d'AGI ou de singularité pour se poursuivre.

Nous allons ici évoquer les différentes études et prévisions sur le sujet, d'abord d'un point de vue quantitatif, puis qualitatif.

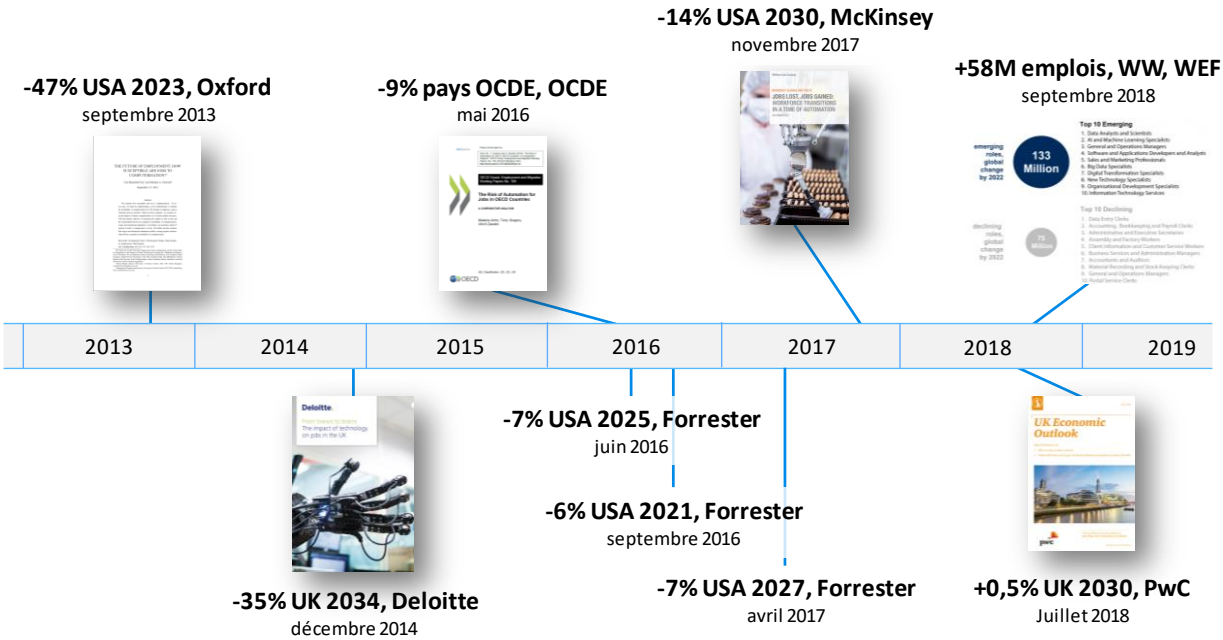
Prévisions

Les prévisions sur la destruction et la création d'emploi liées au déploiement de l'IA dans l'économie sont pléthoriques depuis 2013. On y trouve aussi bien de sombres prophéties sur le rôle même de l'Homme dans l'économie que des prévisions plus optimistes, croyant fermement à la destruction-création de valeur schumpétérienne avec un équilibre positif.

La destruction nette d'emplois liée à l'IA à l'horizon 2023-2025 se situe selon les études entre 6% à 47%, avec des prévisions qui suivent une tendance baissière, la principale prévision de 47% datant de 2013 et celles de 6% à 7% datant de 2016. Ca donne une belle marge d'erreur et de manœuvre !

Cela illustre que les tendances lourdes sur le marché de l'emploi, si elles auront bien lieu, interviendront un peu plus tard. Pour que tel ou tel emploi disparaisse d'ici 5 ans, il faudrait que les technologies correspondantes soient disponibles aujourd'hui compte-tenu de l'inertie du marché, des parties prenantes, des budgets et des déploiements. Si elles ne sont pas encore disponibles, il faudra alors attendre bien plus de 5 ans pour qu'elles aient un impact sur l'emploi ! Un bon nombre de technologies mettent au moins entre 10 et 15 ans à se propager à l'échelle mondiale, surtout si elles nécessitent des infrastructures. Or nombre de prévisions s'appuient sur des technologies qui ne sont pas encore disponibles, même en amont de la R&D.

L'économiste **John Maynard Keynes** se faisait déjà l'écho des risques de pertes d'emploi liées à l'automatisation en 1933, avant même que les ordinateurs fassent leur apparition. Les premières prévisions sur les pertes d'emploi liées à l'IA sont arrivées dans les années 1960. Au démarrage des précédentes révolutions industrielles, les métiers disparus comme les nouveaux métiers ont rarement été bien anticipés. Pour ce qui est du futur, à vrai dire, on n'en sait pas grand chose. Les paramètres à prendre en compte sont tellement importants ! Malgré tout, on se doit de faire des prévisions, même fumeuses, pour anticiper la manière de préparer les générations futures.

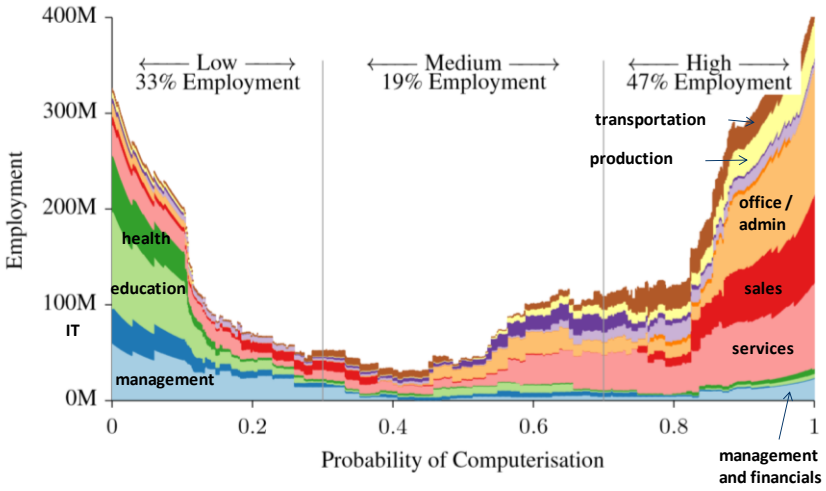


La principale leçon à retenir des prévisions du passé est de conserver un peu d'humilité ! On peut cependant faire quelques hypothèses. Elles sont notamment utiles pour mener certaines politiques publiques, dans l'éducation comme dans les choix de développement infrastructures et de politique industrielle. On sait par exemple qu'il faudra privilégier la formation à des métiers qui ne sont pas répétitifs et où la créativité et l'adaptation jouent un rôle clé.

Je vais faire ici le tour des études qui font ces prévisions sur le futur de l'emploi et en décrire les méthodes et écueils le cas échéant.

- **Septembre 2013** : **Oxford** publie la première grande étude sur l'impact de l'IA sur le futur de l'emploi[789], créée par Carl Benedikt Frey et Michael Osborne. Elle évoque la disparition de 47% des emplois aux USA à l'horizon de 2023. L'étude segmente avec précision les métiers et leurs risques d'être remplacés par des machines. Le calcul du risque s'appuie sur trois formes d'intelligence clés des métiers : l'intelligence motrice (perception et manipulations), l'intelligence créative et l'intelligence sociale. On y constate que la situation est très polarisée : il y a d'un côté des métiers à très faible risque d'automatisation (<20%) comme les fonctions de management, dans la finance, dans le numérique, l'éducation et même la santé, et de l'autre, des métiers à très fort risque d'automatisation (>60%) et surtout dans les services, la vente et l'administratif. La méthode utilisée est mathématique et probabiliste. Nous sommes déjà cinq ans plus tard et les transformations en question sont loin d'avoir eu lieu. Les cycles de l'innovation sont bien plus lents que ce qu'ils prévoyaient. Leur étude sera peut-être valable mais à 15 ou 30 ans d'horizon.

  Cette étude a été souvent critiquée car elle raisonnait au niveau des métiers sans décomposer leurs tâches automatisables ou pas. Nombre d'études qui ont suivi ont repris cette méthodologie en la corrigeant.




THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERISATION?*

Carl Benedikt Frey† and Michael A. Osborne‡

September 17, 2013

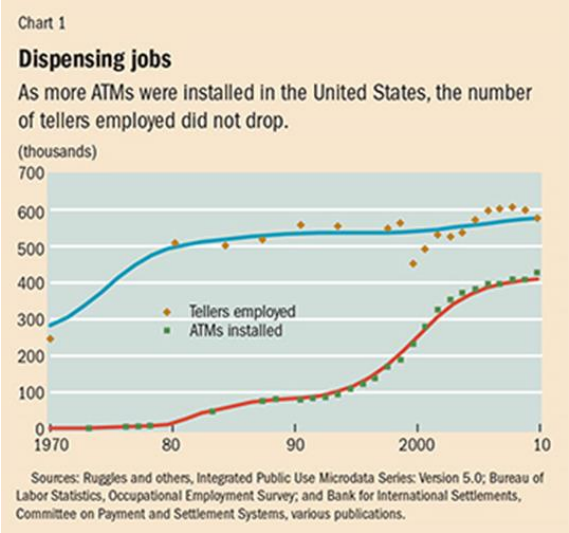
- **Août 2014** : **Pew Research Center** publie une étude qualitative qui recense l'avis de divers spécialistes dont certains estiment que la moitié des emplois sont menacés à l'horizon 2025[790]. Ces experts sont très divisés sur la question ! Le pessimisme provient du risque d'impact rapide de l'automatisation sur les cols blancs avec un risque de déclassification pour un grand nombre, qui seront orientés vers des métiers moins bien payés. Enfin, le système d'éducation ne serait pas en mesure de s'adapter aux nouveaux enjeux. Certains experts sont optimistes car les métiers qui disparaissent sont naturellement remplacés par d'autres, au gré de l'évolution de la demande. La relation avec le travail sera aussi redéfinie de manière plus positive.
- **Octobre 2015** : **Brookings Institution** et Darrell West prévoit des créations d'emplois dans plein de secteurs et des disparitions dans peu de secteurs[791]. C'est une étude qualitative qui se penche sur les politiques publiques à appliquer. Il évoque le revenu minimum, des réformes fiscales diverses, des formations et le développement des activités culturelles !

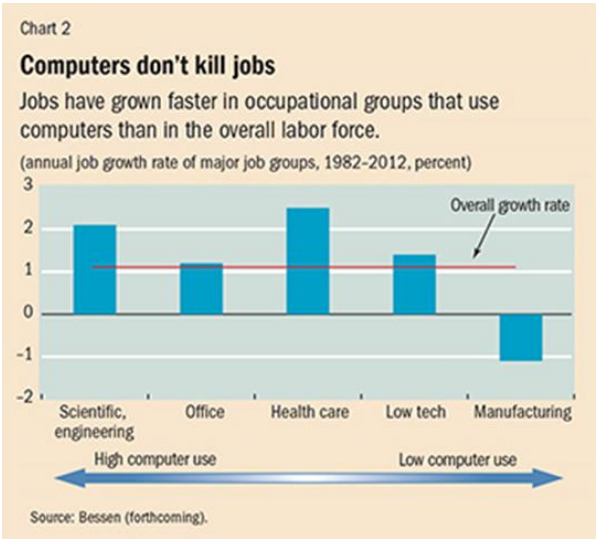
---

[789] Voir The Future of Employment: How susceptible are jobs to computerisation?, septembre 2013 (72 pages).

[790] Voir AI, Robotics and the future of jobs, août 2014 (67 pages).

[791] Voir What happens if robots take the jobs? The impact of emerging technologies on employment and public policy, de Darrell West, octobre 2015 (22 pages).





- **Mars 2015** : **FMI** et James Bessen décrivent l'absence de corrélation entre l'automatisation et les destructions d'emplois. C'est l'origine des schémas *ci-dessus* qui montreraient que les ordinateurs ne sont pas à l'origine de la suppression d'emplois[792].
- **Décembre 2014** : **Deloitte** et **Oxford** publient une étude quantitative qui prévoit le remplacement de 35% des emplois au Royaume-Uni par des robots en 20 ans[793]. Elle s'appuie sur l'Histoire et la destruction d'emplois dans le passé, notamment dans les métiers de secrétariat. Ils intègrent aussi les créations d'emploi et le solde a l'air d'être positif. L'étude est réalisée en collaboration avec Carl Benedikt Frey et Michael Osborne d'Oxford University. Petit détail : cette étude n'a évidemment pas anticipé le Brexit et son impact sur l'économie UK, qui si cela se trouve, sera plus grand que celui de l'automatisation ! C'est le syndrome du cygne noir !
- **Mai 2016** : **OCDE** publie une étude qui anticipe que 9% des emplois sont automatisables dans les pays de l'OCDE[794] mais seulement 6% en Corée du Sud et 12% en Autriche. La France est dans la moyenne à 9%. L'OCDE s'appuie sur l'étude Frey/Osborne de 2013 qu'elle ajuste en corrigeant l'approche de ces derniers par une analyse de l'automatisation des tâches plutôt que des emplois, réduisant par conséquence les destructions d'emplois. L'OCDE n'indique pas d'horizon de temps ni n'évalue les créations d'emplois liées au développement d'innovations et de nouveaux services.
- **Juin 2016** : **Forrester** prévoit que 16% des emplois aux USA vont être automatisés d'ici 2025 que ce soit par de l'IA, des robots ou de l'automatisation classique[795]. Cela sera compensé par la création de 9% d'emplois nouveaux (8,9 millions), générant une perte nette de 7% d'emplois. Ils prévoient que ce sont les emplois de bureau et administratifs qui seront les plus touchés. La création concernera la gestion des robots, les data-scientists et autres techniciens de la robotique et de l'IA ainsi que les curateurs de contenus. Mais ils n'anticipent pas la création d'emplois nouveaux indépendants des impacts technologiques. Ce sont les limites du modèle !

[792] Voir Toil and Technology – Innovative technology is displacing workers to new jobs rather than replacing them entirely de James Bessen, 2015.

[793] Voir From brawn to brains, the impact of technology on jobs in the UK, 2014 (16 pages).

[794] Voir The Risk of Automation for Jobs in OECD Countries, mai 2016 (35 pages).

[795] Voir Robots, AI will replace 7% of US jobs by 2025, Forrester, juin 2016.

- **Septembre 2016** : **Forrester** récidive en prévoyant l'élimination nette de 6% des emplois aux USA[796] d'ici 2021. Ce n'est pas très cohérent avec l'étude précédente publiée trois mois plus tôt, qui évaluait cette perte à 7% en 2025. Pourquoi une perte de seulement 1% sur quatre ans alors que c'est probablement la période pendant laquelle cette perte aurait tendance à s'accélérer ?
- **Mars 2017** : **MIT** et **l'Université de Boston** publient l'étude de Daron Acemoglu et Pascual Restrepo sous l'égide de l'organisme privé National Bureau of Economic Research sur l'impact de la robotisation du le marché de l'emploi US[797] entre 1990 et 2007. Ils déterminent que l'ajout d'un robot dans l'industrie détruit 6,2 postes du fait des répercussions sur l'ensemble de la chaîne de valeur des emplois. Mais cette étude est limitée à l'industrie manufacturière.
- **Avril 2017** : **Forrester** revoit ses prévisions de septembre 2016 à la baisse[798]. Ils anticipent que d'ici 2027 l'automatisation va déplacer 17% des emplois aux USA et en créer 10%. On a donc toujours un solde de 7% mais à une échéance plus lointaine (2027 au lieu de 2025).
- **Mars 2017** : **PwC** publie une étude selon laquelle 38% des emplois US vont être automatisés d'ici 2030 dont 61% dans les métiers de la finance[799]. On revient à des prévisions pessimistes. Les pourcentages équivalents sont de 30% sur UK, 35% en Allemagne et 21% au Japon. Ils tablent pour cela sur une généralisation des véhicules autonomes dès 2020. Ce qui créé évidemment un biais négatif énorme sur l'emploi, notamment pour les conducteurs de camions. Ils anticipent aussi une forte baisse de l'emploi dans la distribution. D'ailleurs, l'IA n'y joue pas un très grand rôle. Les douchettes de self-service dans les hypermarchés et le commerce en ligne est plus en cause.
- **Juin 2017** : **PwC** prévoit que l'IA va permettre d'augmenter le PIB mondial de 14% d'ici 2030, soient $15,7T (trillions de dollars)[800]. Cette croissance serait due pour moitié aux gains de productivité et pour l'autre à l'évolution de la demande. Les gains seront plus forts en Chine (+26% de PIB) et de +14,5% en Amérique du Nord. L'Europe ne génèrerait que 9 à 12% de croissance du PIB et les pays émergents seulement 6% de croissance. Au-delà du fait que la méthode est probablement fantaisiste, on peut espérer que cette croissance ne sera pas indexée comme par le passé sur la consommation d'énergies fossiles. Cela rappelle les études du même genre et ordre de grandeur sur l'IOT qui étaient publiées entre 2013 et 2015. D'autres verront le jour sur la Blockchain et seront également redondantes.
- **Novembre 2017** : **Brookings Institution** analyse la numérisation de 545 métiers couvrant 90% des emplois américains[801]. 70% des emplois créés depuis 2010 requièrent des compétences numériques modérées, et les emplois dont la dimension numérique est plus importante paient mieux.
- **Décembre 2017** : **Gartner** anticipe que l'IA va créer 2,3 millions d'emplois et en éliminer 1,8 millions aux USA d'ici 2020[802] ce qui nous donne un solde positif de 500 000 emplois. Si la croissance économique US suit son cours actuel, il se pourrait que ces prévisions soient justes. Mais pas forcément que l'IA y soit pour quelque chose.

---

[796] Voir Robots will eliminate 6% of all US jobs by 2021, septembre 2016 et The Top Emerging Technologies To Watch: 2017 To 2021, septembre 2016.

[797] Voir Robots and Jobs: Evidence from US Labor Markets de Daron Acemoglu et Pascual Restrepo (91 pages) et Robots and Jobs Evidence from the US Labor Markets des mêmes auteurs, 2016 (69 slides).

[798] Voir Forrester predicts automation will displace 24.7 million jobs and add 14.9 million jobs by 2027, avril 2017.

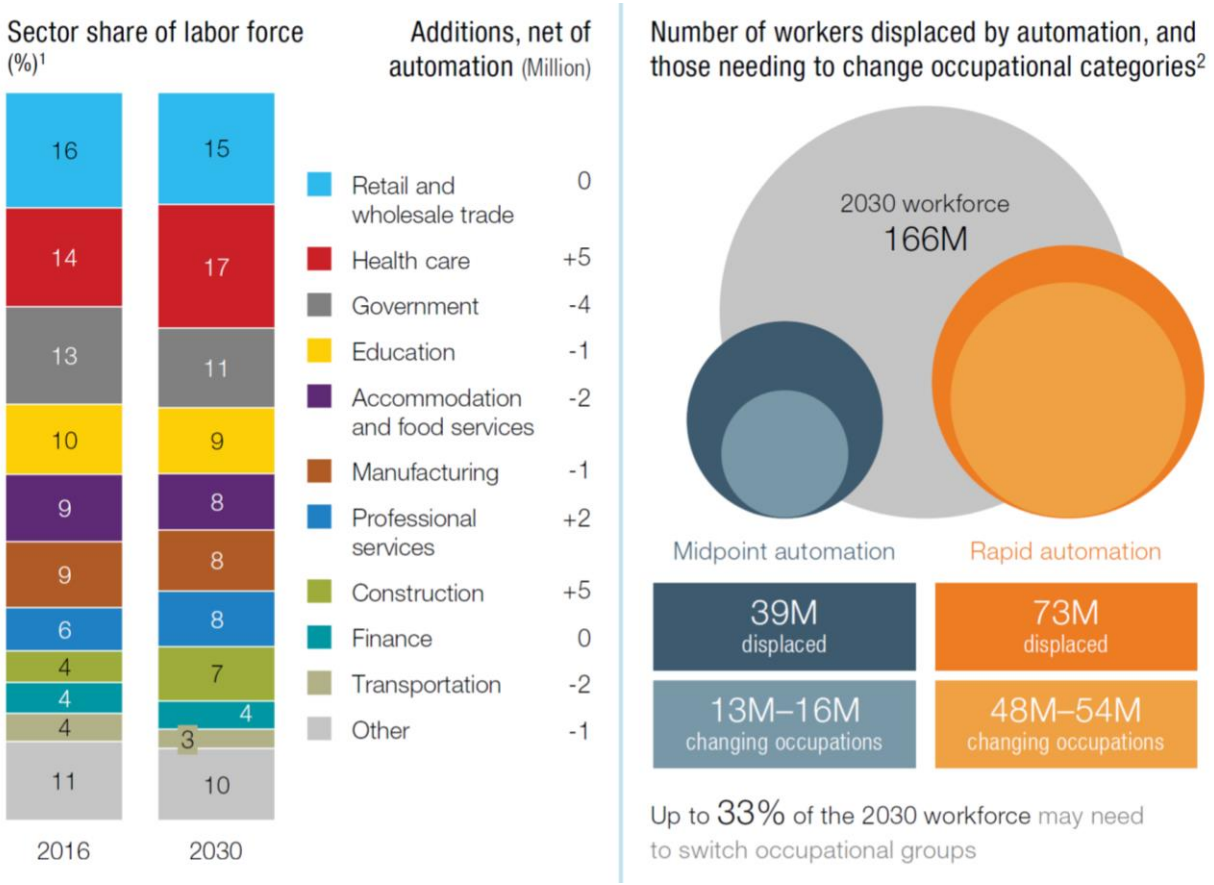
[799] Voir Will robots steal our jobs? The potential impact of automation on the UK and other major economies, mars 2017 (19 pages).

[800] Voir AI to drive GDP gains of $15.7 trillion with productivity, personalisation improvements, juin 2017.

[801] Voir Tech Illiteracy Will Get You Fired Long Before Automation Does, novembre 2017.

[802] Voir Gartner Says By 2020, Artificial Intelligence Will Create More Jobs Than It Eliminates, décembre 2017.

- **Décembre 2017** : **McKinsey** évalue la perte d'emploi à 7 millions aux USA en seulement trois ans. 14% des salariés devraient y changer de métier d'ici 2030 (schéma *ci-dessous*). C'est basé sur le fait qu'ils estiment que 15% des tâches seront automatisées d'ici à 2030, pour 9% en Inde et 24 % aux USA et en Allemagne et 29 % au Japon[803]. McKinsey parle pudiquement de « displaced jobs » quand ils indiquent que des jobs vont être automatisés. Ils ne précisent pas vraiment ce que vont devenir les salariés qui occupent ces jobs qui vont disparaître car les jobs créés ne correspondent pas forcément à leurs compétences.

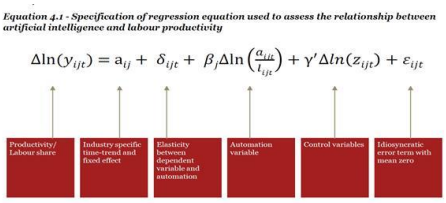


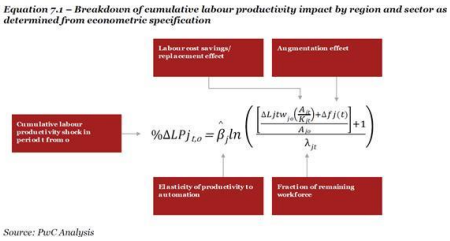
- **Janvier 2018** : **MIT** et l'**Université de Boston** mettent à jour du modèle mathématique de création/destruction d'emploi de Daron Acemoglu du MIT et Pascual Restrepo de la Boston University[804].

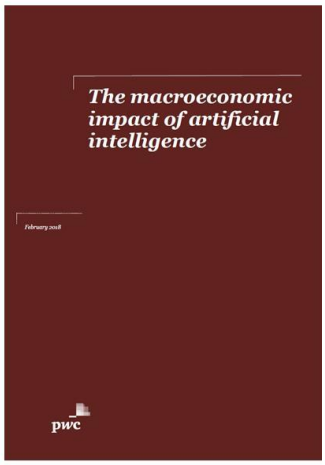
[803] Voir Jobs lost, jobs gained: Workforce transitions in a time of automation, novembre 2017 (160 pages). Voir aussi Jobs lost, jobs gained: Workforce transitions in a time of automation, septembre 2017 (14 slides) qui s'appuie sur l'exemple de l'arrivée des PC.

[804] Voir Artificial Intelligence, Automation and Work de Daron Acemoglu du MIT et Pascual Restrepo de la Boston University, janvier 2018 (42 pages).

- **Février 2018** : **PwC** publie une étude qui fait écho à celle de juin 2017, avec un modèle de prévision très sophistiqué qui couvre UK, Chine, Corée du Sud, Espagne et USA [805]. Ils estiment la croissance du PIB mondial liée à l'IA de 13,8% d'ici 2030. L'IA va impacter 326 millions d'emplois à cet horizon (création + modifications).







- **Mars 2018** : **France Stratégie** publie un rapport sur l'impact de l'IA sur l'emploi en se focalisant sur trois marchés, les transports, la banque, la santé qui représentent 15% du PIB en France[806]. C'est une étude qualitative qui résulte de l'audition d'environ 80 personnes. L'étude aboutit sur quelques recommandations portant sur la formation et sur le lancement d'un « chantier prospectif », le tout s'inscrivant dans le cadre de la mission Villani.

- **Juillet 2018** : **PwC** publie une nouvelle étude sur l'impact de l'IA et de la robotisation sur l'emploi au Royaume Uni[807]. Selon elle, l'IA détruira à peu près autant d'emplois qu'elle en créera à un horizon assez lointain de 12 ans. Au Royaume-Uni, la création/destruction d'emploi portera sur 20% des emplois avec un solde positif de 200 000 emplois sur un total de 7 millions concernés, d'ici 2030, soient 0,5% des emplois. C'est très précis. Et probablement à côté de la plaque à cette échéance là. Ici, le Brexit est censé être pris en compte. L'étude décompose ce solde net d'emplois par région.

Table 4.2: Estimated regional jobs impact of AI based only on variations in industry mix

Region	% of existing jobs (in 2017)			Number of jobs (000s)		
	Creation	Displacement	Net effect	Creation	Displacement	Net effect
London	22.0%	-19.7%	2.3%	1,297	1,159	138
South East	20.6%	-19.7%	0.8%	1,019	978	41
Wales	19.7%	-18.9%	0.7%	302	291	11
Scotland	20.2%	-19.6%	0.5%	558	544	15
South West	19.9%	-19.5%	0.4%	582	571	11
North East	20.0%	-19.8%	0.2%	239	237	2
East of England	20.4%	-20.3%	0.1%	648	646	2
North West	20.4%	-20.4%	0.0%	748	749	-1
West Midlands	20.1%	-20.4%	-0.3%	599	607	-8
Northern Ireland	19.4%	-19.8%	-0.4%	172	176	-4
Yorkshire and the Humber	20.0%	-20.4%	-0.4%	532	544	-12
East Midlands	19.5%	-20.7%	-1.1%	478	505	-27
Total	**20%**	**-20%**	**0%**	**7,176**	**-7,008**	**169**

Source: PwC analysis

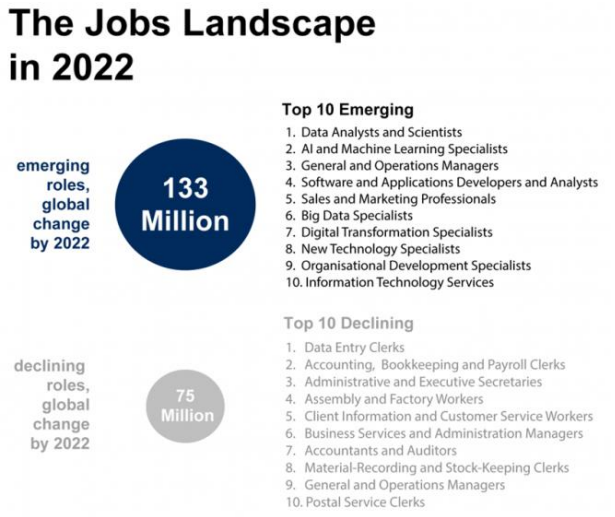
On constate que ce solde est positif dans les grandes métropoles, surtout à Londres, et négatif dans les régions les plus pauvres comme le Yorkshire, les East Midlands et l'Irlande du Nord. L'IA va donc accélérer la concentration de la richesse dans les grandes villes.

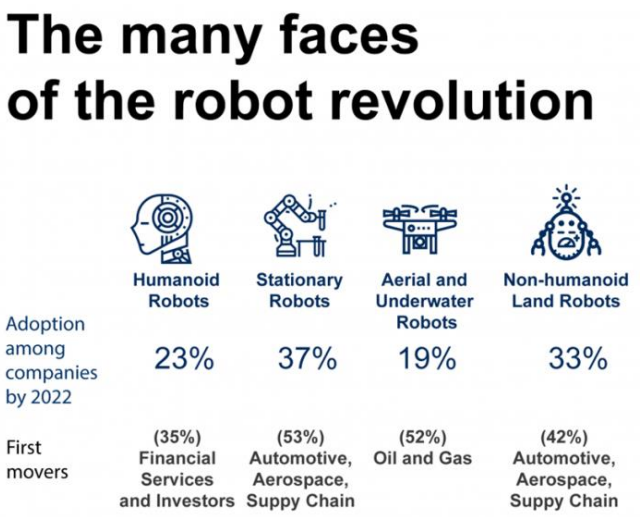
[805] Voir The macroeconomic impact of artificial intelligence, PwC, février 2018 (78 pages).

[806] Voir Rapport Intelligence Artificielle et Travail, France Stratégie, mars 2018 (90 pages).

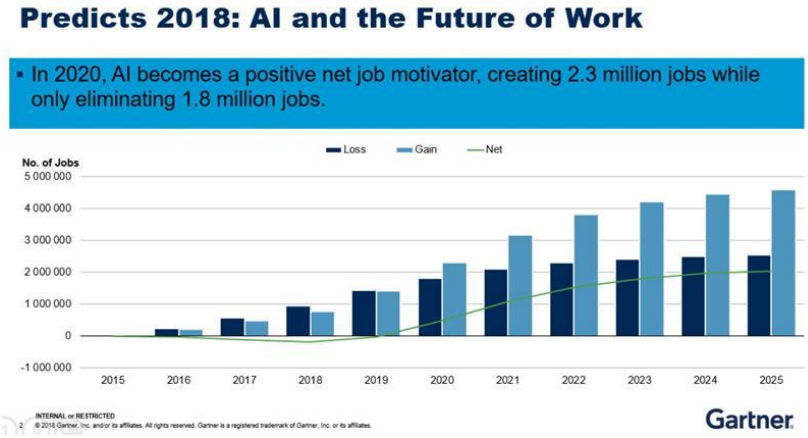
[807] Voir UK Economic Outlook - Prospects for the housing market and the impact of AI on jobs, PwC, juillet 2018 (56 pages).

- **Septembre 2018** : **World Economic Forum** (le WEF, organisation du forum de Davos) évoque la création de 58 millions d'emplois liés à l'IA d'ici 2022 représentant le solde de la création et la destruction de respectivement 122 et 75 millions d'emplois, le tout à l'échelle mondiale[808]. Cela semble bien élevé même si 122 millions ne représentent qu'à peine 1,6% de la population mondiale. Mais ces prophéties se réaliseraient qu'à une condition : que les gens soient bien formés. Ils indiquaient aussi que dans une autre étude, un seul métier a disparu en 60 ans : celui des liftiers d'ascenseurs[809] ! Selon eux, un métier qui est partiellement et non totalement automatisé peut générer de la croissance dans l'emploi car il devient plus accessible à la clientèle. C'est le principe de la commoditisation.





- **Octobre 2018** : le **Gartner Group** prévoit que le solde entre créations et destructions d'emplois lié au déploiement de l'IA commencera à être positif à partir de 2020[810]. Sur quoi est-ce basé ? Sur des enquêtes clients. Ce qui ne veut évidemment rien dire surtout si les données ne sont pas ajustées comme dans les sondages politiques.



  Peu de dirigeants ont le courage d'affirmer dans ces enquêtes qu'ils sont prêts à dégraisser les effectifs grâce à l'automatisation et pour améliorer l'EBITDA de leur entreprise. Cela ferait désordre.

Les prévisions sur l'emploi donnent donc lieu à des affichages de données très exagérées ou avec des trompe l'œil. En voici un exemple que je décortique, « sauce fact-checking ».

---

[808] Voir 2022 Skills Outlook, World Economic Forum, septembre 2018.

[809] Voir One. That's how many careers automation has eliminated in the last 60 years, mars 2017. Ils ont peut être oublié les dactylos qui n'ont peut-être pas entièrement disparues mais sont tout de même des plus rares dans les entreprises. Leur fonction est parfois cumulée avec celle de certaines assistantes ou assistants.

[810] Voir Gartner : 2020, année charnière de l'intégration de l'IA en entreprise, octobre 2018.

Cela commence avec un tweet anodin émis par la talentueuse équipe de la startup française **Kokoroe** qui joue le rôle d'intermédiaire entre personnes souhaitant se former et formateurs. Le tweet indique que 85% des jobs de 2030 n'existent pas encore. Ils seront créés du fait des nouvelles interactions entre les humains et les machines, notamment grâce au déploiement de l'intelligence artificielle dans tous les métiers.

Cela fait beaucoup ! Cela sous-entend donc que seulement 15% des jobs de 2030 seront voisins de ceux d'aujourd'hui. Mais le nombre de jobs différents peut aussi très bien augmenter, réduisant d'autant la proportion des jobs d'aujourd'hui qui n'existerait plus en 2030 !

L'équipe de Kororoe indique la source de son tweet, un article du Figaro qui date de juillet 2017[811]. Cet article fait référence à un rapport de Dell et l'**Institute for the Future**, un think tank californien[812], publié en 2017.

Le chiffre de 85% n'est pas vraiment sourcé dans l'étude de l'**Institute for the Future**. Elle fait référence à un article de Forbes The rise of the freelancer economy de Brian Rashid de 2016 qui fait allusion à l'évolution du marché des freelances selon laquelle en 2020, la moitié de la force de travail US en 2020 en sera constituée.

Les informations de l'article en question de Forbes servent surtout à mettre le lecteur en bouche pour la description d'un service en ligne d'une startup, Tispr, qui joue le rôle de plateforme d'intermédiation entre freelances et employeurs. Donc, avec un biais sur le nombre de freelances.

Pour trouver d'où viennent les chiffres concernant l'emploi des freelances aux USA, il faut chercher dans une l'étude Freelancing in America 2017, de Edelman (68 slides). On y découvre comment les statistiques peuvent être présentées de manière trompeuse. L'étude met en évidence la croissance de la part des freelances dans l'économie. Ils représentaient 35% de l'emploi et 7% de l'économie en 2017. Et en extrapolant à 2027, les freelances seraient aussi nombreux que les salariés.

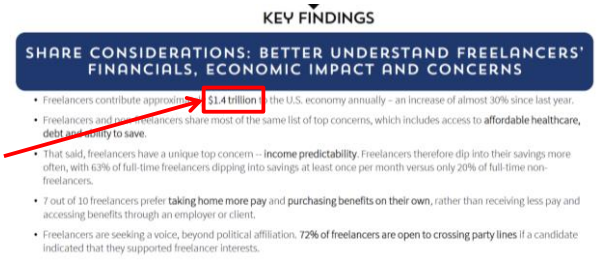
Seulement voilà, il y du double booking ! En 2017, seuls 29% des freelances aux USA le sont entièrement. Les autres sont des freelances à temps partiel à 53% et enfin, 16% sont aussi des salariés. Dans ce dernier cas, leur activité de freelance sert à joindre les deux bouts en fin de mois, surtout pour les travailleurs.

---

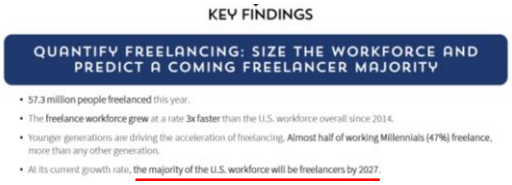
[811] Voir Une étude affirme que 85% des emplois de 2030 n'existent pas aujourd'hui.

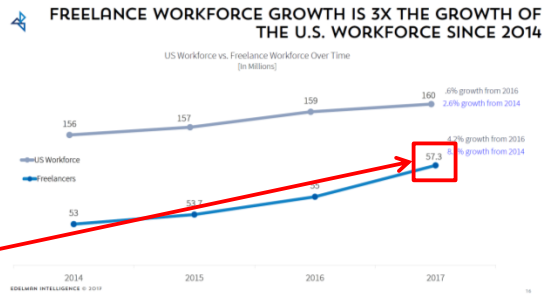
[812] Voir The next era or human machine partnerships, Dell et Institute for the Future, 2017 (23 pages).

On se rend compte aussi que les jeunes ont de plus en plus tendance à devenir freelances lorsqu'ils arrivent sur le marché de l'emploi. Mais les autres classes d'âge sont de moins en moins freelances !



sur $19,39T en 2017 donc **7%** de l'économie US

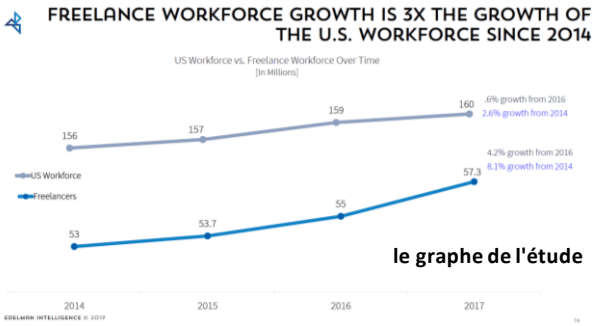




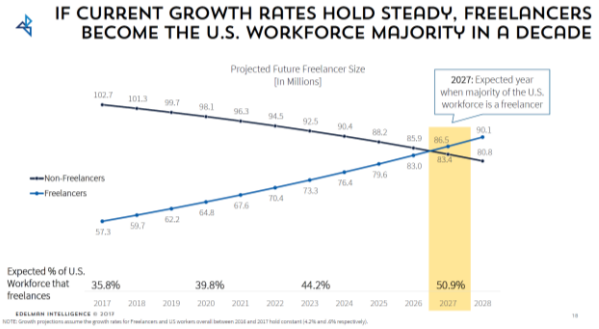
soit **35%.**
l'échelle graphique n'est pas bonne ! Et ces emplois freelances s'additionnent souvent à des emplois salariés

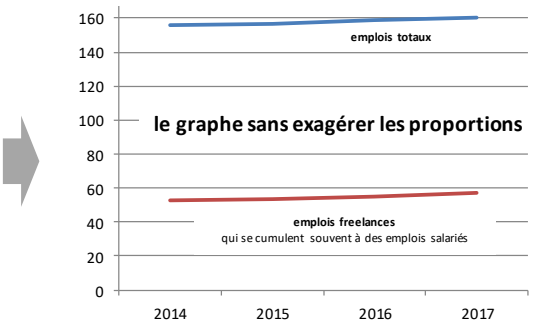
Evidemment, les prévisions à 10 ans ne sont que des extrapolations linéaires de la tendance en cours or on sait que ce genre de méthode est sujette à caution.

Si l'on revient aux 85% d'emplois de 2027 qui n'existent pas encore aujourd'hui, il faudrait préciser s'il s'agit d'emplois entièrement nouveaux ou d'emplois existants dont l'outillage aura évolué et aura nécessité une formation. S'il est certain que la majorité des emplois sera transformée pour les nouveaux outils issus de l'IA et nécessiteront leur lot de formations continues, ces emplois subsisteront tout de même en grande majorité. Qu'il s'agisse des employés de bureau, des ingénieurs, des services publics, des enseignants, des agriculteurs, etc.



le graphe de l'étude





ces courbes font des extrapolations abusives !

si on tirait le trait encore plus loin, il n'y aurait plus d'emplois salariés en 2040 !

ces emplois freelances s'additionnent souvent à des emplois salariés, donc, on ne compare pas des grandeurs homothétiques, il faudrait avoir 3 catégories: salarié, salarié+freelance et freelance.

Seuls certains métiers disparaîtront entièrement ou partiellement du fait des économies réalisées via l'automatisation. Le métier souvent mis en avant est celui de chauffeur de poids lourd. Mais il faudra probablement plus de temps qu'une douzaine d'année pour transformer complètement ce métier. Dans un premier temps, nous aurons surtout des chauffeurs qui rouleront dans des camions semi-autonomes. Au minimum, il faudra les former en conséquence.

Analyses

Nombreuses sont les analyses qui cherchent à segmenter les métiers qui seront transformés de près ou de loin par l'IA. Sont en ligne de mire prioritaire de transformation ou remplacement par les technologies numériques et par l'IA.

De nombreuses précautions doivent être prises lors de prévisions de destruction/remplacement d'emplois résultant de l'automatisation :

- **Etudes et sondages**. Nombre d'études présentées ci-dessus sont le résultat de sondages de dirigeants, pas d'analyse factuelle des technologies à venir et des rythmes prévisibles des innovations. Les enquêtes d'opinion sur le futur sont, comme en politique, à interpréter avec précaution.
- **Ne pas confondre métiers et tâches**. L'IA peut parfois automatiser certaines des dernières mais pas les métiers en entier. C'est le cas d'un cancérologue ou d'un ophtalmologue qui pourra bénéficier de l'apport de systèmes d'imagerie exploitant de la vision artificielle, mais qui auront toujours un rôle d'intégrateur des sources d'information sur le patient, des traitements et de la relation avec le patient dans la durée. L'automatisation amène à une recomposition complexe des métiers difficile à anticiper.
- **Intégrer la dimension macro-économique**. Sont automatisables en priorité les métiers pratiqués de manière homogène sur des marchés larges, qui sont faciles à décrire et automatiser, où les ressources humaines sont soit rares soit trop chères, soit au mauvais endroit, avec des startups financées dans leur secteur et une réglementation favorable aux innovateurs, ce qui n'est pas le cas partout. Cela créé un filtre qui échappe à de nombreuses prévisions. Une analyse sur l'impact de la robotisation sur les emplois devrait porter sur leur structure. Les métiers sont très divers et fragmentés.

  Rien que dans la santé, on trouve des dizaines de types d'emplois et spécialités différentes. Il en va de même dans les services. Les startups s'attaquent en général en priorité à des cibles à la fois faciles et volumineuses, là où l'on peut générer une croissance exponentielle et de belles économies d'échelle au niveau mondial. Les kinésithérapeutes seront-ils remplacés par des robots bipèdes ? Probablement moins rapidement que les conducteurs de camions car ils sont moins nombreux, donc ne présentant pas les mêmes économies d'échelle potentielles ! Et l'automatisation du travail d'un kiné est plus complexe que celle d'un conducteur de camion. Par contre, nombre de métiers sont relativement protégés : ceux qui sont très manuels et difficiles à réaliser par des robots, les métiers créatifs et contenu relationnel[813], ceux dont les tâches ne sont pas répétitives, ceux qui nécessitent des sens très pointus.

  Et puis bien sûr, ceux qui seront créés entre temps. Le monde des loisirs et du futile est assez prolixe de ce point de vue-là. Si l'on observe les nouveaux métiers créés depuis la fin de la seconde guerre mondiale, ils sont dominants dans ces catégories (tourisme, transports, médias, publicité, services divers, boutiques de tatouages, etc.).

[813] Apporter de l'amour en plus de l'IA ! Voir A blueprint for coexistence with artificial intelligence de Kai-Fu Lee, juillet 2017.

- **Pays émergents** : la majorité des prévisions ont aussi une fâcheuse tendance à se focaliser sur la situation aux USA et à ne pas adopter une approche mondiale du problème. Ils n'évoquent pas non plus le cas des fonctionnaires qui sont souvent les derniers à être robotisés car protégés par la lenteur de l'innovation dans les administrations et le manque de courage politique. L'impact de l'IA sur les payés émergents pourrait être encore plus sombre que pour les classes moyennes des pays développés. En effet, si la robotisation se poursuit dans l'industrie, elle supprimera des métiers d'exécution dans les pays émergents et transférera, dans une moindre mesure, de la valeur vers les pays développés, y compris ceux d'Asie (Corée, Japon, Taïwan et une partie de la Chine). L'automatisation des processus administratifs impactera de son côté les métiers de l'offshore, notamment en Inde.
- **Métiers en pénurie de compétences**. Dans le cas des médecins, l'automatisation ne réduira pas forcément l'emploi car le monde manque de médecins et notamment dans de nombreuses spécialités comme en ophtalmologie, en cardiologie ou en diabétologie. Les cancérologues ne sont pas non plus remplacés par IBM Watson. Ce dernier leur permet d'affiner leur diagnostic, leur prescription, et de les rendre plus personnalisés. Le métier de cancérologue est plus menacé par les progrès en médecine prédictive et en immunothérapies que par l'IA. A beaucoup plus long terme, les technologies permettant la prolongation de la vie en bonne santé pourraient cependant réduire le besoin en nombre de médecins, surtout si les maladies dites de longue durée sont éradiquées, cancers, diabètes et maladies neurodégénératives en premier. Certains actes de chirurgie seront aussi de plus en plus réalisés par des robots[814]. Des phénomènes de vases communicants peuvent intervenir. Telle disparition entraîne la création d'emploi dans des secteurs connexes voire entièrement différents des métiers disparus. Pour ce qui est des radiologues, souvent présentés comme étant les plus menacés par l'IA dans les prévisions, ils sont 8736 en France sur un total de 325 870 médecins (**données 2016**). Si l'IA d'analyse des radios améliorait de 50% leur productivité, cela pourrait avoir un impact sur environ 4000 emplois. Donc environ 1,2% du total des métiers dans la santé. A ceci près qu'il faut se soucier du maillage territorial. Les vases communicants peuvent être à l'œuvre. On peut imaginer que le nombre de radiologues ne baisse pas mais que leur métier évolue pour intégrer une partie du travail des manipulateurs radios ! Les migrations de valeur peuvent être sérialisées ! Il faudrait aussi ajouter les 1500 anapats (anatomie pathologiques) en laboratoire dont une partie du travail peut-être assistée avec des techniques d'IA voisines de celles de la radiologie.
- **Création de métiers inconnus aujourd'hui**. Les prévisions s'accrochent trop souvent à notre vision actuelle des métiers, sans anticiper la création de métiers inconnus aujourd'hui, notamment dans le domaine des loisirs et de l'émotionnel avec les métiers de créatifs, les freelances et les métiers manuels dont les services à la personne[815]. C'est aussi l'approche de l'économiste français Nicolas Bouzou, adepte de la prise de recul historique sur les craintes de destruction de l'emploi pour nous rassurer[816]. Les prospectivistes ne sont d'ailleurs pas tous d'accord sur le sort qui sera réservé au métier d'enseignant et de docteur. Certains les voient entièrement remplacés par des robots et de l'IA, d'autres au contraire, non, car la relation avec les élèves et les patients devra rester humaine. C'est une question de perspective sur les aspirations humaines ! Or, si l'automatisation des métiers libère du temps et que le pouvoir d'achat des classes moyennes ne passe pas à la trappe (hypothèse…), alors, elle fera émerger de nouveaux besoins.

---

[814] Voir comme exemple : Surgical robot could sew you up better than a doctor de Jon Figas, avril 2016.

[815] Voir Artificial Intelligence and Robotics and Their Impact on the Workplace produit par l'Iinternational Bar Association Global Employment Institute en avril 2017 (120 pages).

[816] Notamment dans « Le travail est l'avenir de l'Homme », Nicolas Bouzou, 2017.

Accenture PLC publiait en mars 2017 une étude terrain basée sur une enquête auprès de 1000 grandes entreprises qui évaluait la création d'emplois liés à l'automatisation d'autres emplois. Elle en liste quelques-uns, mais néglige les emplois créés dans d'autres domaines affectés par l'automatisation[817].

The Jobs That Artificial Intelligence Will Create (Continued from page 15)

REPRESENTATIVE ROLES CREATED BY AI

Accenture's global study of more than 1,000 large companies identified the emergence of three new categories of uniquely human jobs.

TRAINERS	**Customer-language tone and meaning trainer**	Teaches AI systems to look beyond the literal meaning of a communication by, for example, detecting sarcasm.
	Smart-machine interaction modeler	Models machine behavior after employee behavior so that, for example, an AI system can learn from an accountant's actions how to automatically match payments to invoices.
	Worldview trainer	Trains AI systems to develop a global perspective so that various cultural perspectives are considered when determining, for example, whether an algorithm is "fair."
EXPLAINERS	**Context designer**	Designs smart decisions based on business context, process task, and individual, professional, and cultural factors.
	Transparency analyst	Classifies the different types of opacity (and corresponding effects on the business) of the AI algorithms used and maintains an inventory of that information.
	AI usefulness strategist	Determines whether to deploy AI (versus traditional rules engines and scripts) for specific applications.
SUSTAINERS	**Automation ethicist**	Evaluates the noneconomic impact of smart machines, both the upside and downside.
	Automation economist	Evaluates the cost of poor machine performance.
	Machine relations manager	"Promotes" algorithms that perform well to greater scale in the business and "demotes" algorithms with poor performance.

- **Automatisation sans IA** : il ne sera pas nécessaire d'atteindre un quelconque point de singularité où l'intelligence de la machine dépasserait l'homme pour que les tsunamis de l'emploi se produisent. Ils peuvent intervenir bien avant ! Et pour cause : bien des métiers d'exécution relèvent de tâches très répétitives qui sont sujettes à l'augmentation de l'automatisation dans un premier temps, sans passer par la case de l'AGI, l'intelligence artificielle générale, celle qui remplacerait totalement l'intelligence humaine, puis la dépasserait rapidement par la force démultiplicatrice des machines.

  Les prévisions oublient un autre phénomène induit par le numérique : le transfert du travail non pas seulement vers les machines mais aussi vers les clients, que l'on observe avec les distributeurs automatiques et caisses automatiques, le e-commerce, la SDA (sélection directe à l'arrivée) des centres d'appels, les chatbots (qui peuvent nous rendre aussi rapidement fous que les SDA) tout comme les banques et les assurances en ligne. Comme la valeur économique du temps des gens à faible revenu est faible, elle est absorbée en échange de services en théorie plus rapides[818]. C'est un principe également courant dans l'économie collaborative, qu'elle concerne les professionnels (cas d'Uber, version VTC) ou les particuliers (Blablacar, Aibnb).

- **Démographie** : l'impact des progrès issus de l'IA sur la démographie sont-ils évalués ? Si la durée de vie s'allonge et le confort s'améliore, la démographie pourrait voir sa croissance ralentir, comme c'est le cas au Japon isolationniste depuis quelques décennies. Dans la réalité, elle restera inégale. Les technologies issues de l'IA ne se déploient pas à la même vitesse selon les

[817] Voir The Jobs That Artificial Intelligence Will Create, mars 2017 (5 pages).

[818] Le service n'est pas vraiment plus rapide mais on économise le temps de transport vers un point de service ou de vente.

continents et rien ne dit qu'elles éradiqueront les inégalités sur l'ensemble de la planète, surtout si le moteur de leur déploiement est hautement capitalistique[819].

- **Timing** : on se trompe souvent sur le terme et même la nature des chamboulements. Ils sont généralement surestimés à court et moyen terme et sous-estimés à long terme, mais surtout mal appréhendés dans leur réalité technique et économique[820]. Les études de cas mises en avant dans les ouvrages sur le futur de l'emploi collent souvent à l'actualité marketing du secteur de l'IA. Les livres parus après 2011 commencent presque tous par évoquer la victoire d'IBM Watson dans Jeopardy. A partir de 2013, ils passent aux prescriptions en cancérologie, l'une des applications commerciales de Watson, dont on découvrait les sérieuses limites en 2018. Depuis environ 2011, nous avons droit aux Google Car et autres avancées dans la conduite automatique. En 2016, ce sont les agents conversationnels (chatbots) qui sont devenus d'actualité, du fait de divers lancements comme dans Facebook Messenger.

- **Fact-checking** : en quelques années, les études de cas brandies en trophées peuvent perdre de leur substance. Il a été fait beaucoup de cas de la décision du Taïwanais **Foxconn** en 2011 de déployer un million de robots pour remplacer leurs travailleurs de ses usines en Chine qui demandaient des augmentations de salaire ou se suicidaient. Quatre ans plus tard, seulement 50 000 robots avaient été déployés[821], ce qui ne présage rien de leur capacité à réaliser l'objectif annoncé mais illustre la difficulté à robotiser certains métiers manuels, même répétitifs.

  Dans cette abondante littérature sur le futur de l'emploi, les fondements scientifiques et technologiques des prévisions sont rarement analysés. S'y mêlent allègrement la science-fiction, la science et la fiction[822].

- **Environnement** : les optimistes de l'innovation estiment que, grâce à l'IA, l'Homme sera capable de résoudre tous ces problèmes, presque d'un coup de baguette magique. En exagérant un peu, l'IA est devenue en quelque sorte la solution de sous-traitance ultime des sociétés procrastinatrices et des fainéants : ne nous attaquons pas aux problèmes qui fâchent et attendons que l'IA et la robotique fassent le boulot à notre place ! C'en est presque un éloge du laisser-aller.

---

[819] Dans "The Demographics of Stagnation - Why People Matter for Economic Growth" de Ruchir Sharma dans Foreign Affairs, march-april 2016 selon qui la robotisation arrive à temps pour acompagner la baisse de la démographie dans les pays développés. Le Japon est un bon exemple : il cherche à produire des robots pour prendre en charge les personnes âgées car il n'y a pas assez de jeunes (ou d'immigrés) pour s'en occuper. Il y a moins de jeunes qui arrivent sur le marché du travail avec un effet retard de 18-22 ans sur cette baisse démographique. L'article ne le dit pas, mais la France a la particularité d'avoir une meilleure natalité qu'ailleurs en Europe. Mais ne la transforme pas pour autant en croissance et en emplois contrairement à de nombreux autres pays. Donc, la France est potentiellement plus vulnérable que d'autres pays à la robotisation des métiers.

[820] Ainsi, dans Les robots veulent déjà nous piquer notre job d'Emmanuel Ghesquier qui commente une étude d'un certain Moshe Vardi de l'Université Rice du Texas, il est indiqué que l' "*on a pu voir avec les robots Pepper que certains robots pouvaient donner des conseils de gastronomie ou d'œnologie dans les supermarchés Carrefour ou qu'une boutique de téléphonie allait fonctionner à 100% avec des employés robotisés au Japon*". L'auteur qui relaie cela n'a pas du voir Pepper à l'œuvre car, au stade actuel de son développement, il est encore plus que brouillon ! J'avais même pu le constayer en 2014 dans une boutique Softbank dans le quartier Omotesando[820] où ils commençaient à être déployés. Et ce n'est pas mieux dans toutes les démonstrations que l'on peut voir de ce robot dans différents salons professionnels. En y regardant de près, l'étude en question est un article publié dans The Conversation, Are robots taking our jobs. Il a bien du mal à faire le tri dans les évolutions de l'emploi aux USA entre ce qui provient de l'automatisation, de la globalisation et de la concurrence asiatique dans l'industrie manufacturière et même indienne, dans les emplois concernant les services informatiques. L'emploi a surtout migré géographiquement. Les emplois de perdus dans l'industrie aux USA et en Europe se sont retrouvés en Asie. C'est le "monde plat" de Thomas Friedman.

[821] Voir http://www.generation-nt.com/foxconn-foxbot-robot-assemblage-humain-actualite-1914702.html.

[822] Dans le top de l'exagération technique, nous avons par exemple **Tomorrowland** de Steven Kotler (2015), qui prédit monts et merveilles singularistes allant de l'intelligence artificielle générale (AGI) autoréplicable jusqu'au téléchargement des cerveaux dans un ordinateur : "*Yet it is worth noting that Moore's Law states that computers double in power* ***every twelve months*** *[…]. Biotechnology, meanwhile, the field where mind uploading most squarely sits, is currently progressing at* ***five times the speed of Moore's Law****. […] people alive today will live long enough to see their selves stored in silicon and thus, by extension, see themselves live forever.*". Nous avons donc une loi de Moore deux fois plus rapide dans les processeurs que dans la vraie vie (12 vs 24 mois) et des "biotechnologies" qui évoluent cinq fois plus rapidement que la loi de Moore, alors que cette vitesse ne concerne que le cas particulier de l'évolution du coût du séquençage de l'ADN, observée sur la période courte 2007-2011. Evolution qui s'est plutôt calmée les 5 années suivantes [822]!

Les ressources à notre disposition sont-elles infinies[823] ? L'IA nous sauvera-t-elle assez rapidement d'un éventuel réchauffement planétaire accéléré dans les 30 ans qui viennent ? Les priorités de l'humanité pourraient en tout cas être sérieusement chamboulées.

- **Révoltes** : la *Ludditisation* des métiers n'est généralement pas évoquée dans les prévisions, du nom des Luddites qui résistèrent au début du 19ième siècle contre le développement des machines à tisser au Royaume-Uni. Tandis que la Reine Elisabeth I avait refusé l'octroi d'un brevet à William Lee en 1589, après son invention de la machine à tisser les bas, craignant de générer du chômage chez les ouvriers textiles, le gouvernement de sa Majesté avait décidé d'envoyer la troupe contre les ouvriers récalcitrants au progrès, entre 1806 et 1811. Un gouvernement élu par un parlement dominé par des entrepreneurs ! Quelles forces pourraient résister à l'automatisation des métiers ? Certains métiers ont-ils une meilleure capacité de résistance que d'autres, notamment par la voie de la règlementation ? Nous avons peu d'exemples résilients dans le temps !

Métiers

Segmentons donc les métiers qui seront impactés de près ou de loin par l'automatisation et par l'IA avec les métiers du passé qui disparaissent déjà, les métiers en train d'être automatisés qui vont disparaitre ou être transformés, les métiers menacés, les métiers qui vont être créés et les métiers protégés.

Métiers du passé

L'automatisation n'est pas nouvelle dans l'Histoire humaine. Elle a commencé il y a des millénaires avec l'agriculture, l'usage de bêtes de traits puis de tracteurs. Elle bat son plein depuis le 19e siècle et la première révolution industrielle. Elle s'est amplifiée avec l'avènement des outils numériques.

Les ouvriers de lignes d'assemblage ont ddéjà remplacés par des robots et le seront de plus en plus, surtout au gré du *in-shoring*, le rapatriement de la fabrication dans les pays occidentaux.

Les caissiers sont remplacés en partie par des automates de self-service et donc, par le travail gratuit des clients.

Les centres d'appels de taxis sont remplacés par des applications mobiles quand ce n'est pas la commande vocale. Aussi bien chez les sociétés de VTC à la Uber tout comme chez G7, le leader du marché traditionnel.

Métiers menacés

Les métiers en train d'être partiellement automatisés ne disparaîtront pas mais l'amélioration de leur productivité pourra réduire les effectifs[824]. Cela concerne surtout les cols blancs et de nombreux métiers de services, notamment dans les professions libérales administratives et dans la finance qui est de plus en plus automatisée[825]. Dès lors qu'une tâche est répétitive ou qu'elle nécessite une prise de décision selon des règles assez simples, l'automatisation peut prendre le relai. C'est encore plus vrai si le marché est très concurrentiel et faiblement régulé.

---

[823] Deux ouvrages intéressants traitent assez bien de ces questions : **The beginning of infinity** de David Deutsch, qui défend un point de vue selon lequel l'infini et l'innovation sont intimement liés et qu'il ne faut pas de mettre des barrières à notre capacité d'innovation. Et puis **The infinite resource** de Ramez Naam qui fait un bilan circonstancié des défis qui se présentent pour gérer les ressources en apparence limitées de la planète côté énergie, agriculture et matières première. Il équilibre bien ces difficultés et les progrès techniques à venir qui permettront de les contourner.

[824] Voir L'impact de la révolution digitale sur l'emploi - Top 5 des métiers en voie de disparition de Erwann Tison, Institut Sapiens, 2018 (25 pages).

[825] Voir How my research in AI put my dad out of a job And what we are doing with the French government to prevent other people from losing theirs de Rand Hindi, fondateur de Snips.ai. Mai 2017.

L'IA automatise facilement les fonctions de reconnaissance d'image, d'où un impact sur un métier spécifique dans la santé : les radiologues. Cela ne veut pas dire qu'ils disparaîtront mais que leur productivité augmentant, ils pourront traiter plus de patients.

Cela concerne aussi les métiers de l'offshore comme les sous-traitants en Inde de processus d'entreprises qui pourraient être automatisés par les techniques de Robotic Process Automation que nous avions évoquées dans la rubrique sur ce vertical[826].

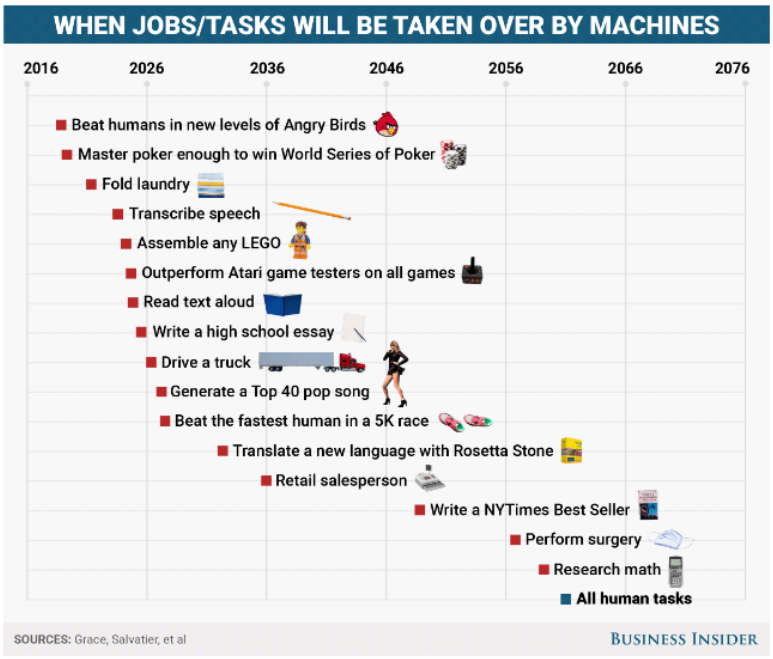
On trouve dans ce créneau de nombreux métiers dans les banques et les assurances, dans la comptabilité, dans les fonctions de secrétariat. Côté travail manuel, la manutention pourrait progessivement passer à la moulinette de la robotisation dans les grands entrepots.

Métiers qui seront automatisés plus tard

Totalement ou partiellement comme dans la santé, le management, l'audit et même dans la recherche[827]. L'activité de conducteur professionnel sera partiellement automatisée du fait de la généralisation des véhicules autonomes.

Mais on aura toujours besoin d'eux pour les chargements et déchargements et pour les trajets dans les zones denses. Il se pourrait bien que l'on ait besoin d'autant de conducteurs qu'aujourd'hui, comme pour les pilotes d'avions. Par contre, ils seront moins fatigués par leur travail du fait de l'automatisation de la conduite sur les longs trajets sur autoroutes.

Certains prospectivistes vont jusqu'à prévoir que le métier de développeur va lui-même être automatisé, ce à quoi je ne souscris pas du tout[828]. Le schéma ci-dessous illustre les tâches qui vont être progressivement automatisées. Cela menacera donc les métiers qui les exécutent[829]. Mais attention : tout ce qui dépasse la quinzaine d'année est totalement spéculatif. Il est très difficile de prédire les évolutions technologiques à de telles échéances. Il n'y a qu'à voir ce que l'on prédisait dans les années 1970 ou 1980 pour 2000, avec des taxis volants et autres fantasmes. Il faut se méfier des prévisions extrêmes[830].



---

[826] Voir Introduction to Robotic Process Automation, a primer, de l'IRPA, Institute for Robotic Process Automation, 2015 (35 pages).

[827] Voir Les prochains paradigmes d'exploration scientifique seront peuplés d'Intelligences Artificielles, d'abord assistantes, elles deviendront collaboratrices, puis chercheuses d'Aymeric Poulain Maubant, octobre 2016.

[828] Ce métier va évoluer comme il a évolué sur 50 ans, avec des transformations profondes, des différences accrues entre les développeurs de solutions utilisateurs assemblant des briques préexistantes et ceux qui les créent. Au même titre qu'aujourd'hui, la compétence pour faire évoluer le noyau de Linux n'est pas la même que pour créer un site en Wordpress avec des templates et des plugins. Il n'y a pas d'automatisation du métier, mais plutôt une stratification entre couches « hautes » et « basses » requérant des niveaux de compétences différentes.

[829] Voir This is when robots will start beating humans at every task, juin 2017, qui fait référence à When Will AI Exceed Human Performance? Evidence from AI Experts, mai 2018 (21 pages).

[830] Comme celles de Yuval Harari, cité dans Pour l'auteur de Sapiens, l'intelligence artificielle va détruire la plupart des emplois, Challenges, octobre 2017.

Métiers protégés

Différentes études inventorient les métiers qui ne sont pas menacés par l'automatisation, par l'IA ou la robotique[831]. En voici quelques-uns :

- Les **métiers créatifs**, qui feront encore appel à des talents humains, même s'ils sont outillés par de l'IA pour étendre la palette d'expression. Dans tous les cas d'IA créative dont nous sommes abreuvés et qui font appel aux réseaux génératifs, le choix des outils et de la création reste l'affaire du créatif.
- Les **techniciens de maintenance**, même si leur activité est planifiée par des IA et de la maintenance prédictive et quand bien même on peut voir des démonstrations impressionnantes de robots de Schaft, maintenant chez Softbank Robotics. Cela concerne aussi les artisans dans le BTP, la plomberie et l'électricité.
- Les **métiers de services** qui touchent au corps : coiffeurs, manucures, kinésithérapeutes, etc. Ils ne sont pas prêts d'être remplacés par des robots. D'ailleurs, certains de ces métiers sont parfois plus menacés par la démocratisation d'outils exploitables par les consommateurs eux-mêmes.
- Les **professions de santé** sont en général faiblement menacées, surtout celles d'infirmiers/ières, les chirurgients (même s'ils peuvent utiliser des robots), les médecins généralistes (qui vont pouvoir utiliser des outils de diagnostics plus nombreux), et de nombreux spécialistes. Si l'IA va certainement automatiser certaines tâches, elles portent sur des métiers qui sont en pénurie, comme les ophtalmologues. L'amélioration de leur productivité ne va donc pas les faire disparaître.
- Les **travailleurs sociaux** pour lesquels le contact humain est clé[832].
- Les **enseignants** sont souvent mis en avant comme une profession protégée car essentielle. Et ce, malgré les nombreuses idées d'usage de l'IA dans l'éducation pour l'enseignant à distance et pour l'accompagnement des élèves. Nous avons vu pourquoi dans la rubrique de cet ebook sur l'éducation : la praticité de nombreuses solutions d'IA n'est pas évidente du fait de la grande fragmentation des activités de formation.

Métiers qui vont être créés.

Reste à étudier les métiers qui vont se développer voir être créés du fait de l'IA ou indépendamment de l'IA.

- Il y a bien entendu tous les **ingénieurs et développeurs en IA** et les métiers de support du numérique qui sont associés. Une bonne partie de l'écosystème numérique va basculer dans l'IA comme il a basculé dans l'Internet. L'IA sera rapidement une commodité qui perfusera dans tous les secteurs du numérique (startups, éditeurs de logiciels, constructeurs, sociétés de services, conseil, formation).
- Les **métiers dans les loisirs** sachant que de nouveaux loisirs seront créés qui feront appel ou pas à l'IA. Dans ceux qui y feront appel, on peut évidemment penser aux applications de la réalité virtuelle et augmentée.
- Les métiers des **énergies renouvelables**. On le dit depuis longtemps et ce n'est pas lié à l'IA.

Revue de lectures

Voici quelques ouvrages de référence sur le sujet du futur du travail.

---

[831] Comme These are the few jobs that robots won't take from us, de Michael Grothaus, août 2018.

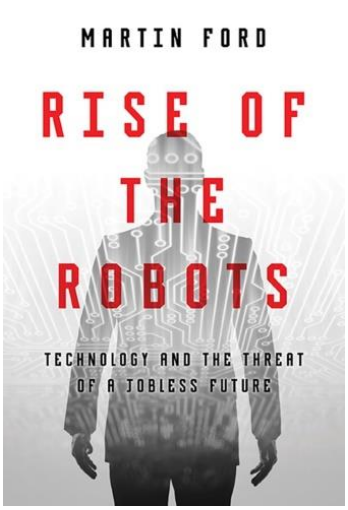
[832] Voir Yes, AI may take some jobs – but it could also mean more men doing care work, septembre 2018.

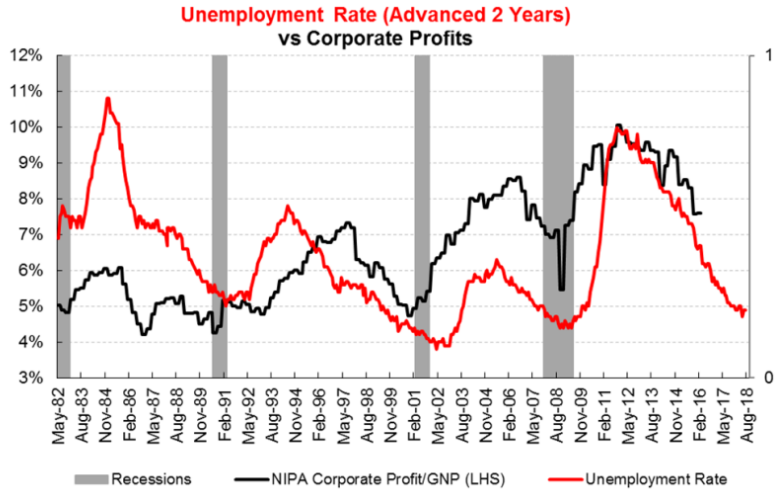
Rise of the robots and the threat of a jobless future (2016) de Martin Ford est un ouvrage bien documenté qui évoque un bon nombre des mécanismes macro-économiques des précédentes révolutions et crises industrielles, et de ce qui pourrait advenir dans le futur.

Sa thèse principale est que les révolutions numériques passées et à venir contribuent à réduire l'emploi dans les classes moyennes et à favoriser d'un côté l'émergence d'emplois de bas niveaux mal payés et de l'autre d'emplois de haut niveau bien payés. C'était déjà anticipé dans le rapport **Triple Revolution** produit en 1964[833] pour l'administration de Lyndon B. Johnson. Ses auteurs s'alarmaient déjà sur risques les risques de l'automatisation, mettant en avant la difficulté de remplacer les emplois supprimés par la modernisation à un rythme suffisamment rapide. Il était très en avance sur son temps, alors que l'informatique n'en était encore qu'à ses balbutiements. Juste avant la sortie du mythique mainframe IBM 360, en 1965, c'est dire !

Aux USA, les 5% des foyers les plus aisées représentaient 27% de la consommation en 1992 et 38% en 2012. Les 80% les moins aisés sont passés de 47% à 39% dans le même temps.

Après la crise de 2008, le top 5% avait augmenté ses dépenses de 17% et le reste n'avait fait que rester au niveau de 2008. D'ou l'émergence de business comme Tesla qui cible, pour l'instant, surtout les 5% les plus riches. Les nouvelles entreprises issues du numérique sont automatisées dès le départ et ont moins de salariés. Elles profitent à plein de la productivité issue du numérique. Les exemples un peu éculés et trop généralisant de Whatsapp et Instagram sont mis en avant pour illustrer le point. On nous bassine un peu trop avec les $16B de "valeur" de Whatsapp générés par 55 employés, alors que lorsqu'elle a été acquise par Facebook, cette société n'avait quasiment pas de revenus.





Contrairement à l'après-guerre, les gains de productivité des deux dernières décennies sont allés non pas dans l'augmentation des salaires mais dans la baisse des prix, dans les salaires de métiers techniques qualifiés, et le capital s'orientant vers le financement des nouveaux investissements technologiques. Les technologies sont devenues un facteur d'inégalité au profit des technologues et des détenteurs de capital, tout du moins aux USA. La "finance" réalloue aussi les profits au bénéfice des plus riches. Plus un pays a un système financier développé, plus grandes seraient les inégalités.

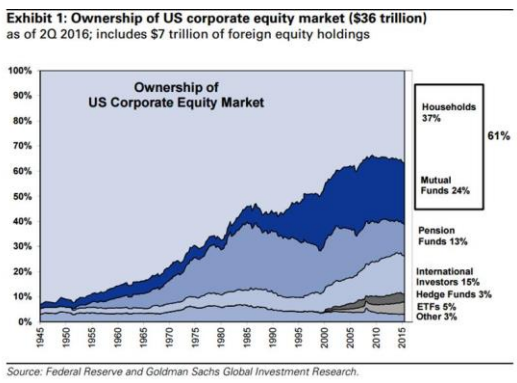
Les profits des grandes entreprises ont augmenté sur 15 ans en proportion du PIB comme indiqué dans le schéma *ci-dessus* (source), qui correspond aux données US. Cette réallocation concerne 2,5% du PIB. Je me suis demandé où allaient ces profits.

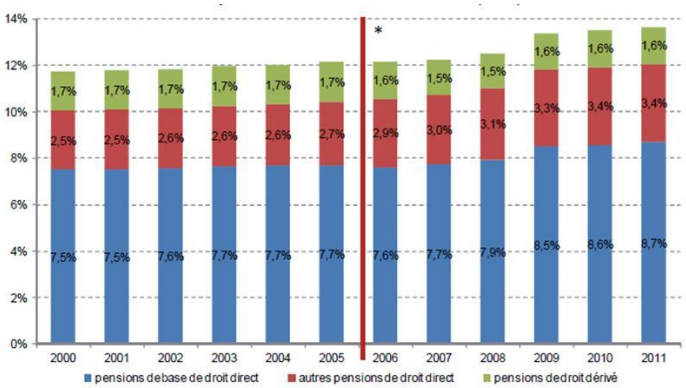
Un tiers alimente les fonds de pension. Un autre tiers va dans les foyers, et probablement avec des inégalités fortes de revenu. Le reste va pour moitié chez des investisseurs internationaux, certains,

[833] Voir The Triple Revolution, Libération, 1964.

aussi pour alimenter des fonds de retraite. Les méchants fonds spéculatifs (hedge funds) ne représentent que 4% de l'actionnariat des entreprises américaines [834]!

Et si l'explication était donc toute simple : plus la population vieillit, plus les systèmes de retraites par capitalisation ont besoin de financement, donc de profits des grandes entreprises !





En France, le régime général des retraites a vu son poids dans le PIB évoluer de 11,2% en 1990 à 13,8% en 2008, soient 2,6% de progression. Coïncidence ? Ne serait-ce pas finalement une solution différente au même problème ? A savoir, augmenter les charges sociales et taxes pour financer une retraite par répartition en lieu et place d'une augmentation du profit des grandes entreprises qui rémunèrent un système de retraite par capitalisation ? C'est probablement à moitié vrai et à moitié faux car les profits des grandes entreprises françaises ont aussi augmenté dans la même période. Mais comme les actions du CAC40 sont détenues par des investisseurs étrangers, il se trouve qu'ils alimentent aussi les systèmes de retraite de pays étrangers, notamment anglo-saxons qui en sont friands !

Le paradoxe est que la pénurie de compétences qualifiées ralentit ce phénomène de concentration de la valeur sur les plus riches ! Si on y pourvoyait plus rapidement, cela détruirait encore plus de jobs mal payés, et bien plus que de jobs bien payés de créés. La limitation des visas de travail pour les cadres qualifiés étrangers imposée par le congrès US créerait une inertie souhaitable pour protéger les emplois non qualifiés. En même temps, elle favorise l'offshore de métiers qualifiés en plus des métiers faiblement qualifiés qui sont déjà externalisés à l'étranger. Et la politique immigatoire encore plus restrictive de Donald Trump ne va pas arranger cela.

Autre point intéressant, l'auteur fait état des écueils des MOOCs, présentés comme la solution miracle pour l'enseignement. Deux études menées par l'Université de Pennsylvanie en 2013 et qu'il ne faudrait pas forcément généraliser montrent que les résultats d'étudiants ayant suivi des MOOCs étaient moins bons que ceux d'étudiants passant par des méthodes traditionnels. Il ne faut certainement pas jeter le bébé du MOOC avec l'eau du bain de ces études. Les méthodes mixant MOOCs et enseignement IRL (in real life) sont probablement à favoriser.

L'ouvrage de Martin Ford met aussi en avant des opinions divergentes sur l'avenir de l'IA. L'expert en sciences cognitives **Gary Marcus** trouve que les performances récentes de l'IA sont survendues. Pour **Noam Chomsky**, qui s'est penché sur les sciences cognitives pendant 60 ans, on est encore à des millénaires de la création de machines intelligentes comme l'homme et que la singularité reste du domaine de la science fiction. Même opinion pour le psychologue cognitiviste **Steven Pinker**, le biologiste **P. Z. Myers** et même pour **Gordon Moore**. Il évoque aussi l'histoire de la National Nanotechnology Initiative lancée en 2000, qui survendait l'idée de créer des nano-machines au niveau des atomes et s'est ensuite rabattue sur des objectifs plus raisonnables.

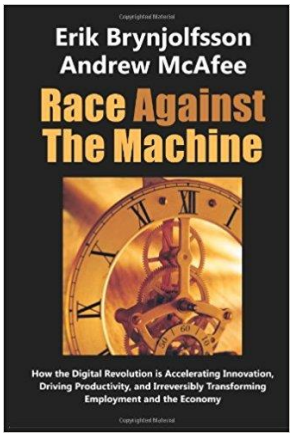
[834] Voici la **source** du schéma correspondant.

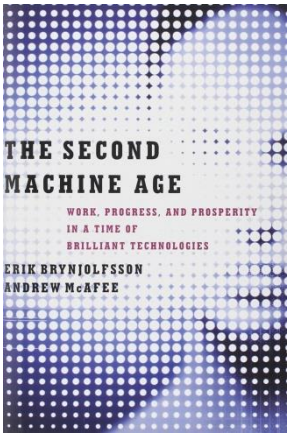
Martin Ford évoque l'intérêt du revenu minimum qui est souvent présenté comme la solution pour traiter le problème de la disparition trop rapide d'emplois liés à la robotisation[835]. C'est une sorte d'Etat providence générique poussé à l'extrême quand il n'est plus en mesure de créer les conditions d'une activité pour tous.

Ces débats font rage avant même que la richesse permettant de le financer ne soit créée et que de nouveaux métiers soient automatisés. La Finlande est parfois mise en avant comme validant le principe alors que le revenu minimum n'y était qu'à l'état expérimental entre 2017 et 2018 et n'a d'ailleurs pas été reconduit en avril 2018[836].

Les questions clés sont nombreuses. Quel est le niveau de ce revenu minimum ? Est-il là juste pour simplifier les systèmes existants de redistribution ? Comment est-il financé s'il est plus élevé ? Comment est-il différentié en fonction de la situation des foyers ? Comment évite-t-il de décourager les gens de travailler là où cela reste nécessaire ? Quel serait son impact si mis en place dans des pays et pas dans d'autres ? Quel impact sur les flux migratoires qui créent déjà une pression certaine ? Il existera toujours des inégalités marquées entre pays, en plus de celles qui existent entre milieux sociaux. Ce débat a démarré il y a plus de 11 millénaires avec les débuts de l'agriculture. Il s'est poursuivi avec toutes les autres révolutions technologiques et industrielles suivantes et n'est pas prêt de se terminer.

Dans **Race against the machine** (2012) et **The Second Machines Age**" (2014), Erik Brynjolfsson et Andrew Mcafee font les mêmes constats que le livre précédent sur la concentration de la richesse sur les 5% les plus aisés. Ils rappellent qui, si l'on considère aujourd'hui encore que les anciennes révolutions industrielles ont créé tant d'emplois, c'est parce que l'on a enlevé de l'équation les chevaux et autres bêtes de somme qui ont perdu leur utilité et ont disparu au passage, ou bien, ont été transformés en chair à steaks.





Ils étaient ce que sont aujourd'hui les travailleurs à bas salaire dont l'activité est en voie d'être automatisée, modulo les steaks. Le bilan écologique est aussi bien connu : c'est la terre qui a payé le prix de la croissance humaine !

Ils décrivent le scénario de l'offshore qui pourrait menacer l'emploi dans les pays à faible cout de main d'œuvre : les métiers délocalisés étaient les plus codifiables et donc, automatisables en priorité lorsque la technologie le permettra. Cela protège pour une part les pays occidentaux. A ceci près que les métiers codifiables non délocalisables pour des raisons physiques sont aussi automatisables. A contrario, le développement des robots réduit l'intérêt des délocalisations dans l'industrie. Il permet en théorie une relocalisation des usines, et la création d'emplois locaux de production, d'installation et de maintenance de robots ainsi que dans la supply chain.

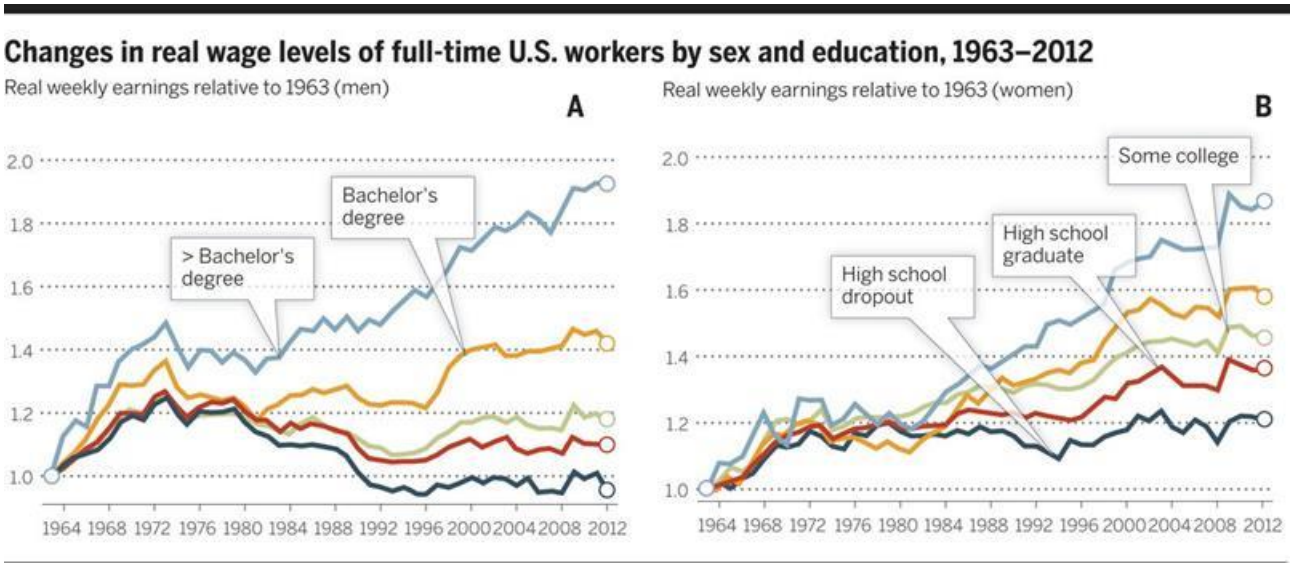
---

[835] Ce revenu minimum (UBI : universal basic income en anglais) est tout de même proposé par Mark Zuckerberg, Richard Branson et Elon Musk. Voir Tax robots and Universal Basic Income de Ben Bloch, juillet 2018.

[836] Voir Finland to end basic income trial after two years, avril 2018. Ce revenu minimum couvrait 2 000 personnes sans emploi âgées de 25 à 58 ans, avec un versement mensuel de 560€. La période de test n'était cependant pas suffisante pour en constater les effets, notamment en termes de réinsertion. Le gouvernement souhaitait revoir les conditions d'attribution et les associer à une formation professionnelle ou à l'acceptation d'un travail. Dans Pourquoi le revenu universel sera un interdit aussi fort que l'inceste de Laurent Alexandre, janvier 2018, insiste bien sur ce point : le RBI ne doit pas remplacer l'impérieuse nécessité de la formation pour adapter les sans emplois à l'économie de la connaissance. En France, le RSA est de 551€ pour une personne seule. On est donc au même niveau et on peut donc considérer que le revenu minimum est déjà en place en France. En fait, l'appréciation porte surtout sur le montant. Celui-ci est dans tous les cas en dessous du seuil de pauvreté. Et pour cause, celui-ci est souvent défini en proportion du revenu médian dans un pays. Il y a donc toujours une partie de la population qui est sous ce seuil.

Le scénario des auteurs met en avant les mêmes gagnants et perdants : les personnes à haut niveau de qualification vs les personnes faiblement qualifiées, les entreprises superstars à croissance exponentielle et les autres, et enfin le capital contre le travail. Il s'appuie sur le fait que, ces dernières décennies, les salaires ont déjà augmenté pour les personnes les plus qualifiées et baissé pour les moins qualifiées (source des graphes *ci-dessous*).

On pourrait ajouter à cette analyse la possibilité d'un ajustement de la population mondiale en fonction des glissements de valeur provoqués par la robotisation. Quelle serait l'influence de la robotisation sur la natalité ? Et surtout de la prolongation de la durée de la vie, sans même parler de vie éternelle. Plus la longévité augmente, comme au Japon, plus la natalité est en baisse. A court et moyen terme, cela résout le problème de l'emploi par le vide. Mais une société vieillissante peut enclencher son déclin inexorable. L'impact d'un éventuel revenu de base ne serait pas neutre. Avec lui, la démographie n'irait plus naturellement à la baisse.



Les deux auteurs, qui sont de la MIT Sloan School of Management, proposent un plan d'action en quatre points qui s'inspire en partie des propositions du **rapport Triple Revolution** de 1964 :

- Investir dans l'**éducation**, en payant mieux les enseignants, en les rendant responsables, et attirer aux USA les immigrants qualifiés. Côté cursus, ils recommandent d'investir dans la créativité, dans l'identification de tendances et dans la communication complexe. Ils font remarquer que l'homme plus la machine sont plus puissants qu'une machine seule[837]. Donc, associer la créativité et la maîtrise de l'usage des technologies reste une belle protection. Ils considèrent que tous les métiers qui requièrent à la fois de la créativité et une forte sensibilité motrice ne sont pas prêts d'être automatisés (cuisiniers, jardiniers, réparateurs, dentistes). Les auteurs font aussi preuve de bon sens en rappelant que notre imagination est limitée pour prédire les emplois du futur. On n'anticipe pas assez la nature des problèmes existants et à venir qui vont générer leurs propres métiers.
- Développer l'**entrepreneuriat** : l'enseigner comme une compétence dans l'ensemble de l'enseignement et pas seulement dans les meilleures business schools, réduire les règlementations qui ralentissent la création d'entreprise, et créer un visa pour les entrepreneurs. Ce visa s'est retrouvé dans l'initiative "Startup Visa Act" lancée en 2011 par l'administration Obama mais qui n'est toujours pas validée par le Congrès US… et qui n'est pas prêt de l'être. Ils re-

[837] Dans **Human+Machine**, Paul Daugherty et H. James Wilson insistent sur le couplage homme plus machine et non pas le remplacement de l'homme par la machine. L'ouvrage est bien commenté par Yves Caseaux dans Réinventer les processus et les applications avec l'Intelligence Artificielle, Yves Caseaux, septembre 2018.

commandent aussi d'encourager les innovations d'organisation et du travail collaboratif pour exploiter ce qu'il reste d'utilisable du temps et des compétences des gens inoccupés.

- Développer l'**investissement** dans l'innovation, la recherche et les infrastructures, notamment dans les télécommunications. Un grand classique des pays modernes comme des pays émergents.
- Côté **lois et fiscalité**, ne pas alourdir la législation du travail. Rendre les embauches plus attractives que la robotisation des métiers au niveau des charges sociales et taxes, ce qui rappelle une bonne partie de la politique de l'emploi en France, qui ne nous réussit pas si bien. Ne pas réguler les nouvelles activités. Réduire les subventions aux emprunts immobiliers et les réallouer à l'éducation et à la recherche. La propriété immobilière a tendance à réduire la mobilité géographique. Réduire les subventions directes et indirectes aux services financiers. Réformer le système des brevets et réduire la durée d'application du copyright.

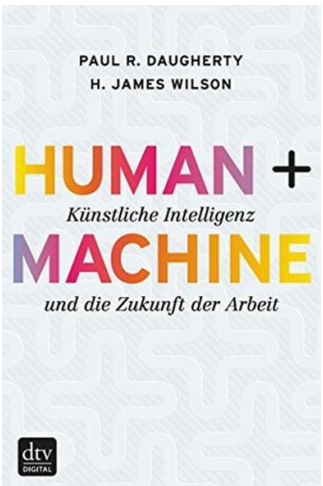
  Enfin, ils ne recommandent pas de créer une allocation universelle mais plutôt un crédit d'impôt pour les bas revenus (negative income tax) dans la lignée d'une proposition de Thomas Paine qui date de 1797 au Royaume-Uni. Pourquoi valoriser le travail ? Parce que, quelle que soit sa nature, en plus de pourvoir à nos besoins, le travail traite deux nuisances : l'ennui et le vice (Voltaire), sans compter les couches hautes de la pyramide des motivations de Maslow.

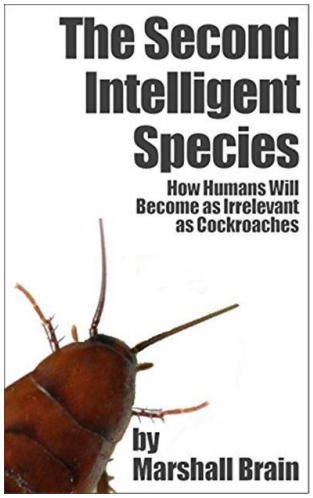
C'en est presque un plan "à la Macron" : favorisons l'entrepreneuriat et tous les problèmes sociétaux se régleront d'eux-mêmes. Un peu trop classique !

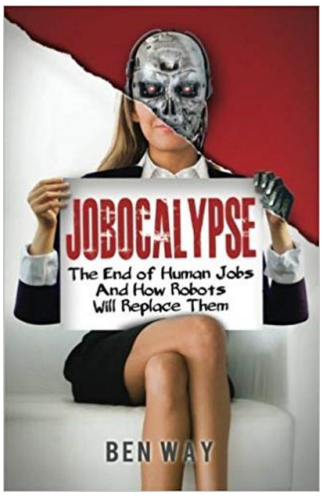
The Second Intelligent Species: How Humans Will Become as Irrelevant as Cockroaches (2015), de Marshall Brain, grossit le trait en annonçant que les scientifiques sont en train de créer une seconde espèce intelligente, les robots et l'IA, qui va nous dépasser et supprimer la majorité des emplois. Les premiers touchés seront les millions de camionneurs, les vendeurs dans la distribution de détail, dans les fast foods et le BTP. C'est un darwinisme technologique provoqué par l'Homme, qui se fait dépasser par ses propres créations.

Le reste est de la non-science-fiction, tablant sur une intelligence artificielle qui régulerait les comportements humains néfastes, comme ceux qui affectent l'environnement. Les emplois non qualifiés disparaitraient à la fin des années 2030, ce qui semble un peu rapide au vu de la progression de la robotique.

Au passage, l'auteur fournit une explication du fameux paradoxe de Fermi selon lequel il est bizarre qu'aucune civilisation extraterrestre ne nous ait approchés à ce jour. "Officiellement", diraient les conspirationnistes. L'IA développée par ces civilisations serait comme la nôtre : une fois qu'elle serait satisfaite par ses réalisations et par l'équilibre ainsi généré, elle n'aurait pas besoin d'explorer le reste de l'univers.

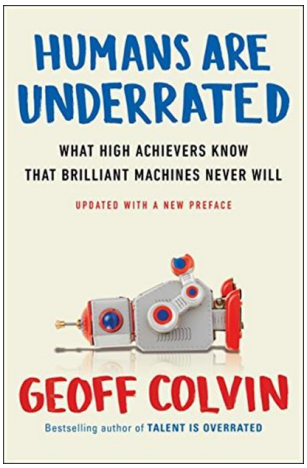






Jobocalypse (2013) de Ben Way, que je n'ai pas lu (désolé…), part du principe que nous sommes *déjà* envahis par les robots et que la disparition d'emplois liée à l'automatisation est une histoire ancienne. Il anticipe que même les métiers les plus qualifiés seront remplacés par des robots car ils s'autoalimenteront. Les scénarios envisagés vont de révolutions provoquées par les sans-emplois à des initiatives gouvernementales de formation massive les concernant. On dira que l'on préférera le second scénario au premier même si c'est un peu court !

Quand au Rapport **Global Catastrophic Risks 2016** de la Global Challenges Foundation, il intègre l'IA dans les risques systémiques que l'humanité et la planète pourraient rencontrer, au même niveau que les conséquences du réchauffement climatique et les pandémies naturelles ou artificielles. Les risques évoqués ne concernent cependant pas les conséquences sur l'emploi mais plutôt la perte de contrôle de l'IA par l'Homme. Cette évaluation est mise à jour dans l'édition 2017 du rapport[838].



Enfin, **Humans Are Underrated: What high achievers know that brilliant machines never will** (2015) de Geoff Colvin met en avant de son côté l'opportunité de remettre au gout du jour les qualités humaines dans les métiers : l'empathie, l'intuition, la créativité, l'humour, la sensibilité et les relations sociales. Une manière de différentier clairement les machines et l'homme.

C'est aussi l'approche proposée par Bruno Teboul dans **Robotariat – critique de la robotisation de la société**, paru au printemps 2017, qui associe philosophie, économie, prospective et humanisme pour envisager un monde équilibré où l'IA et les data sciences ne sont pas mises au pilori, et sont utilisées pour faire avancer la société, et où la place de l'homme et de la nature sont préservées. Il y propose un revenu universel de reconversion à vie financé par les entreprises qui automatisent, de taxer le trading à haute fréquence et de développer une vie plus écologique[839].

Elle était reprise par Dov Seidman dans **Harvard Business Review** en 2014[840] pour qui les métiers du futur sont les métiers qui ont du cœur ! C'est une belle conclusion, même si frisant quelque peu l'utopie[841].

---

[838] Voir Global Catastrophic Risk 2016 (55 pages) et Global Catastrophic Risk 2017 (67 pages).

[839] Cet ouvrage très instructif permet de découvrir ou de redécouvrir de nombreux auteurs clés de ces différents domaines. Il décrit avec recul des phénomènes récents comme l'uberisation. Et il partage quelques défauts avec ce document : des parties de deux à trois pages qui survolent de nombreuses thématiques, un panorama peu-être un peu trop large et pas assez profond, et un jargon pas forcément accessible comme ces « Prolégomènes à une herméneutique des NBIC » (en langage courant, on dirait peut-être « Prélude à une interprétation des NBIC ». !

[840] Voir From the Knowledge Economy to the Human Economy de Dov Seidman, novembre 2014.

[841] Voir aussi cette alerte sur la tendance des outils numériques à réduire les véritables interactions humaines Eliminating the human et à limiter notre prise de risques : L'impossible voyage connecté, ou comment le numérique a étouffé le sentiment d'aventure.

Parades

Maintenant que le problème est posé, comment ne pas être remplacé par de robots et de l'intelligence artificielle ? Après l'uberisation qui intermédie les métiers de service, la robotisation peut-elle automatiser ces mêmes métiers ? La robotisation serait-elle la forme ultime d'ubérisation ?

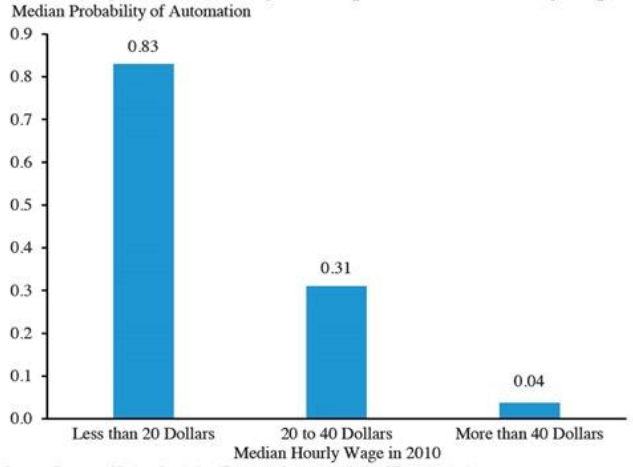
Quelques pistes sont bien connues et déjà citées dans les livres évoqués ci-dessus : choisir des métiers où la créativité, l'initiative, les émotions, l'empathie et l'humanité sont importantes et adopter les nouvelles technologies qui rendent plus productif. Et ne tombons pas dans le panneau des annonces tonitruantes d'IA créatives ! Ce sont des outils pour les créatifs, pas des IA qui se passent de créatifs.

Comme avec toute nouvelle technologie, de nouvelles formes de créativité humaine verront le jour. Les outils de l'IA permettent aux créatifs de tout poil de se poser de nouvelles questions. Un scientifique peut ou pourra explorer la connaissance et l'état de l'art plus facilement. Un chercheur pourra faire des hypothèses et les vérifier plus facilement. Un ingénieur pourra simuler encore plus aisément ses créations. Un urbaniste pourra évaluer l'impact d'un projet. Un marketeur pourra faire de même avec des hypothèses produit et marché. L'IA permettra de créer de nouveaux outils de compéhension de l'existant et de simulation de nouveaux projets dans tous les domaines.

L'abondance des données exploitable par les IA ne fait pas tout ! Il faut savoir se poser les bonnes questions pour les exploiter !

C'est une constante dans l'innovation : dans presque tous les métiers, l'automatisation et la robotisation ne sont jamais totales. Elle nécessite une supervision humaine. Il faut donc s'approprier les outils de cette supervision, voire les créer soi-même ! Donc, de préférence, maîtriser à la fois des métiers traditionnels et les technologies numériques qui peuvent les transformer.

Figure 5-15
Probability of Automation by an Occupation's Median Hourly Wage



Source: Bureau of Labor Statistics; Frey and Osborne (2013); CEA calculations.

Malheureusement, les sciences et technologies n'attirent pas tant que cela les jeunes et notamment en France, comme une enquête mondiale réalisée par **Randstad**[842] le montre.

A contrario, il faudra de préférence éviter les métiers répétitifs, routiniers ou à faible degré de créativité et d'initiative et simples d'un point de vue moteur. Ce sont ceux qui présenteraient le plus grand risque d'automatisation.

Le schéma *ci-dessus* issu du **Rapport Economique du Président US 2016**[843] rappelle que les métiers à bas salaire, donc en général à faible qualification, sont les plus menacés par l'automatisation.

Dans **The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution** publié en janvier 2016 par le World Economic Forum[844], les auteurs prévoient que les deux tiers des enfants en école primaire d'aujourd'hui exerceront un métier qui n'existe pas encore. Ils y vont un peu fort car l'échéance n'est pas si lointaine.

[842] Voir Un Français sur quatre conscient d'être remplacé par un robot, avril 2016.

[843] Voir Economic Report of The President 2016, février 2016.

[844] Voir The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution, janvier 2016.

Ils prévoient que 7,1 millions d'emplois administratifs disparaitront d'ici 2020, et que seulement 2 millions d'emplois seront créés dans les technologies (aux USA). Par contre, des emplois devraient être créés pour combler une partie du trou dans l'énergie, les nano-biologies et le divertissement, et ceux des commerciaux subsisteront. Et oui, les emplois de l'avenir seraient surtout ceux dont le contenu émotionnel sera le plus dense, comme expliqué dans **Les 10 compétences clés du monde de demain**[845].

D'un point de vue stratégique, on peut intuitivement privilégier l'enseignement supérieur, la recherche et l'entrepreneuriat dans les domaines scientifiques et technologiques qui génèrent ces automatisations.

Il vaut mieux créer ou adopter les outils de l'automatisation que de n'en subir que les effets, comme décrit dans **How To Avoid Being Replaced By A Robot** paru dans Fast Company en avril 2016[846].

On pourra aussi favoriser les enseignements pas trop spécialisés et assez diversifiés. Et enfin, ne pas oublier d'exceller dans ce qui fait de nous des Hommes, de belles machines biologiques douées d'émotions.

Politique par l'IA

Les métiers de la sphère marchande ne sont pas les seuls à être impactés par l'IA. Celui de politique l'est tout autant, même s'il n'a pas vocation à être une activité à temps complet dans toutes les démocraties.

Le cerveau fonctionne très souvent par analogies et la connaissance de l'Histoire influe sur les décisions des politiques, sauf lorsqu'ils ne connaissent pas du tout l'Histoire comme Donald Trump.

L'IA n'utilise pas encore massivement le raisonnement par analogies. Il répond surtout en fouillant dans de vastes dépôts de connaissances et pour croiser quelques informations structurées. Mais sait-on…

Politique fiction

Est-ce qu'une IA pourrait indiquer : si tu envahis tel pays dans telle et telle circonstance, voici ce qui a le plus de chances de se produire en suivant les leçons de l'histoire connue. Et voici ce qui permettrait d'éviter le pire ! Une IA pourrait-elle guider un exécutif dans les choix de politiques à la fois rationnelles et irrationnelles ?

On apprend souvent du passé pour (mieux ?) décider du futur. Comment des décisions politiques complexes influencent la sphère économique qui agit à la fois de manière rationnelle et irrationnelle aux événements ?

Autre difficulté à surmonter pour l'IA, mais pas insurmontable : comment tenir compte d'un adversaire qui agit de manière non rationnelle ? La plupart des algorithmes d'IA sont conçus de manière rationnelle. Exemple : comment réagir quand l'une des parties agit de manière irrationnelle, tel un Saddam Hussein en 1990/1991, voire lorsque les deux parties sont irrationnelles avec ce même Saddam Hussein et Georges W. Bush en 2003 ?

Je m'étais aussi demandé en 2013[847], pour les 50 ans de l'assassinat de JFK, si un système de type Watson ne pourrait pas un jour analyser toute la littérature sur le sujet et pondre une synthèse voire résoudre l'énigme qui est bien plus complexe qu'une simple théorie du complot style 9/11 ou sur les chemtrails. L'analyse des faits et mystères de l'histoire pourrait probablement gagner de ce genre de système. Mais l'intérêt économique de la chose est plutôt marginal !

---

[845] Voir Les 10 compétences clés du monde de demain, mars 2016.

[846] Voir How to avoid being replaced by a robot ?, avril 2016.

[847] Voir Les technologies et l'assassinat de JFK, novembre 2013.

Est-ce que les organisations politiques et des Etats peuvent se faire eux-mêmes disrupter par de l'IA ? Evitons l'expression "uberiser" qui est à la fois trop précise et trop vague. Il y a bien l'initiative **Watson for President** mais elle est un peu légère car construite comme une opération de communication d'IBM[848]. Elle visait un peu à la manière de Coluche en 1980, de faire élire Watson comme nouveau président américain en 2016. En indiquant que cela permettrait à la Maison Blanche de prendre des décisions rationnelles.

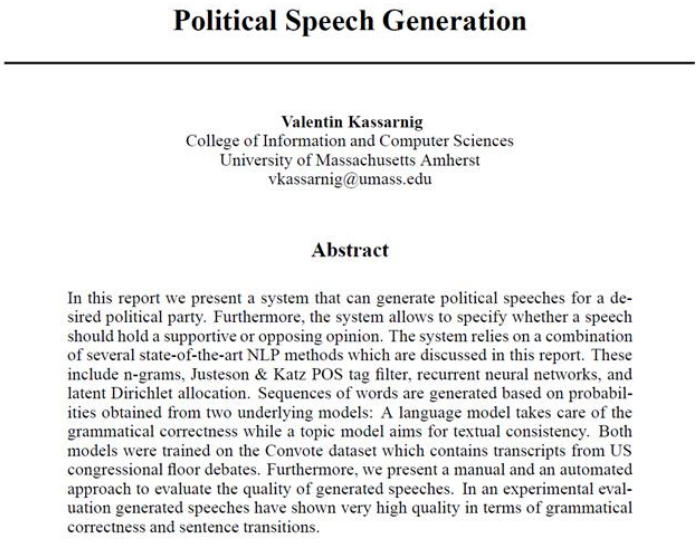
C'est confondre un peu rapidement l'outil de la prise de décision (POTUS) et l'outil d'aide à la prise de décision (Watson et/ou le staff du Président et son administration).

Un président fait déjà appel à de nombreux experts pour prendre ses décisions, en particulier dans la diplomatie, les négociations internationales et le pilotage du bras armé des USA. Il a aussi besoin de pas mal d'aide et de tacticiens pour faire voter des lois par le congrès qui est souvent récalcitrant, même lorsqu'il est du même bord que lui. On l'a vu pour l'Affordable Care Act (Obamacare) lors du premier mandat de Barack Obama. Mais avec l'élection de Trump, on peut songer à l'avantage qu'il y aurait eu à élire une IA !

La première question à se poser sur l'usage de l'IA concerne les élections dans les démocraties. Les dernières grandes élections, notamment américaines, ont montré la force à la fois des réseaux sociaux et de la propagation d'idées véhiculant du rêve (Obama, Sanders) ou des peurs et angoisses, et les fameuses fake news (Trump). L'élection de Trump a montré comment la manipulation des opinions pouvait faire basculer de peu une élection[849].

Que ferait l'IA pour améliorer un tel processus ? Elle collecterait des volumes gigantesques d'informations ouvertes sur ce qui se dit et s'écrit, sur ce que font les électeurs, sur leurs réactions à des discours antérieurs, sur les analyses biométriques (de la captation de pouls avec une montre, des mouvements oculaires avec des capteurs de Tobii, de l'EEG pour la mesure de l'activité cérébrale, etc), sur l'économie ou sur les médias.

Elle les analyserait alors au point de permettre la création de programme politiques appliquant soit la **démagogie ultime** (celle qui fait gagner les élections mais qui est inapplicable ou qui, si appliquée, mène à une catastrophe) soit la **démagogie utile** (celle qui fait à la fois gagner les élections et aller dans un chemin non catastrophique et responsable). Le tout en étant conforme à une idéologie de base d'un parti politique donné, avec son système de valeur (partage, social, économie, croissance, environnement, fiscalité, justice, école, selon les cas). Voilà de beaux défis d'optimisation sous contraintes !



Political Speech Generation

Valentin Kassarnig
College of Information and Computer Sciences
University of Massachusetts Amherst
vkassarnig@umass.edu

Abstract

In this report we present a system that can generate political speeches for a desired political party. Furthermore, the system allows to specify whether a speech should hold a supportive or opposing opinion. The system relies on a combination of several state-of-the-art NLP methods which are discussed in this report. These include n-grams, Justeson & Katz POS tag filter, recurrent neural networks, and latent Dirichlet allocation. Sequences of words are generated based on probabilities obtained from two underlying models: A language model takes care of the grammatical correctness while a topic model aims for textual consistency. Both models were trained on the Convote dataset which contains transcripts from US congressional floor debates. Furthermore, we present a manual and an automated approach to evaluate the quality of generated speeches. In an experimental evaluation generated speeches have shown very high quality in terms of grammatical correctness and sentence transitions.

Des tentatives de ce genre ont déjà été vaguement lancées. Valentin Kassarnig, chercheur à l'Université Amherst du Massachusetts, a présenté début 2016 un premier **générateur de discours politique** basé sur de l'IA[850], et qui dépasse les générateurs de pipeau déjà bien connus.

---

[848] Voir http://watson2016.com/.

[849] Hillary Clinton a devancé Donald Trump en vote populaire de 2,8 millions de voix mais a perdu le college des grands électeurs pour 78 000 électeurs dans quatre *swing states*, qui avaient fait l'objet d'un ciblage particulier de fake news dans les réseaux sociaux. J'avais fait une analyse chiffrée de cette élection dans L'origine et les conséquences de l'élection de Donald Trump, novembre 2016.

[850] Voir Political Speech Generation de Valentin Kassarnig, 2016 (15 pages).

Le résultat reste assez rustique et focalisé sur le langage, pas sur la construction d'un programme politique qui se tienne. La solution est même diffusée en open source[851] ! Malheureusement, en politique plus qu'ailleurs, l'adage selon lequel le contraire de l'IA est la bêtise naturelle s'applique parfaitement. Cette dernière est même plutôt efficace électoralement !

Après les élections se pose la question de la gestion. Est-ce que l'IA permettrait de préparer des choix censés mis ensuite dans les mains d'électeurs dans le cadre de démocraties plus participatives ? Est-ce que l'IA permettrait de bâtir des politiques économiques dignes de ce nom ? Est-ce l'IA permet d'intégrer les complexes relations sociales dans la société ?

D'anticiper les réactions des citoyens aux nouvelles lois et réglementations, notamment fiscales ? Est-ce qu'elle permettra de gérer les conflits ? Est-ce qu'elle pourrait permettre d'accélérer la justice ? D'éviter les erreurs judiciaires ? De réformer les systèmes de santé au fil de l'eau des progrès technologiques ? Je n'en sais rien. Il n'y a pas beaucoup de chercheurs qui planchent sur ces questions ! Certains indiquent toutefois qu'une IA impliquée dans le processus apporterait un peu de rationalité et serait capable de prendre des décisions non basées sur le côté obscur des émotions[852].

Les systèmes d'aide à la décision politique pourraient-ils faire appel à de l'IA intensive ? Y compris lorsqu'il s'agit d'évaluer la position et l'attitude des autres parties prenantes, des agents économiques ou des chefs d'Etat ? Est-ce qu'une IA permettrait à un POTUS[853] de gérer de manière optimale la relation conflictuelle avec Vladimir Poutine, les bras de fer avec les Chinois, ou de résoudre pacifiquement les divers conflits du Moyen-Orient ? Ou à un successeur de François Hollande de se dépatouiller de la situation en France ?

On a bien vu des films de Science Fiction mettant en scène des personnages liés à l'IA comme dans "Her" et "Ex Machina", mais pas encore dans la politique fiction. Ca ne saurait tarder vue l'imagination débridée des scénaristes ! Un « House of Cards » avec un « Special Assistant to the President » qui soit une IA à commande vocale ferait son effet et nous éloignerait de la présidence-réalité actuelle !

On en est encore loin. Ce qui démontre par l'absurde que l'AGI (Artificial General Intelligence) n'est pas pour tout de suite. Mais gare à vos fesses les politiques ! La démocratie participative pourrait prendre un visage inattendu !

Politique réalité

L'IA est en fait déjà utilisée en politique dans quelques contextes sporadiques.

Le machine learning a été utilisé pour la segmentation et le ciblage d'audiences clés dans diverses élections par l'anglais **Cambridge Analytica**[854].

---

[851] Voir https://github.com/valentin012/conspeech.

[852] Voir Should Politicians be Replaced by Artificial Intelligence? Interview with Mark Waser, 2015.

[853] POTUS = President Of The United States.

[854] L'origine, l'histoire et les méthodes de Cambridge Analytica sont bien documentées dans la présentation Uses and abuses of AI in election campaigns de Alistair Knott (https://ai-and-society.wiki.otago.ac.nz/images/0/0f/Ai-and-elections.pdf). La société a été créée en 2013 et financée par l'investisseur Bob Mercer, un ultra-conservateur, cofondateur du site d'information Breitbart. Cambridge Analytica avait comme VP un certain Steve Bannon, passé pendant 7 mois à la Maison Blanche comme conseiller spécial de Donald Trump. Pour la présidentielle 2016, la société exploitait diverses sources de données acquises, notamment auprès d'Acxiom et Experian. Cela leur a permis de constituer une base nominative de plus de 200 millions d'américains adultes avec 5000 données associées. Ils ont même fait des expériences d'A/B Testing sur des messages TV s'appuyant sur leurs données de profiling. Voir aussi Cambridge Analytica, the shady data firm that might be a key Trump-Russia link, explained de Sean Illing, octobre 2017. J'écrivais ceci en octobre 2017. Depuis, le scandale de Cambridge Analytica a éclaté au printemps 2018 aux USA du fait de la capture de données privées sur Facebook de dizaines de millions d'utilisateurs par Cambridge Analytica en 2015 et 2016. Cela a entraîné une commission enquête du Sénat US puis la fausse fermeture de Cambridge Analytica devenu Emerdata. Voir The power players behind Cambridge Analytica have set up a mysterious new data company, mars 2018 et Cambridge Analytica dismantled for good? Nope: It just changed its name to Emerdata, mai 2018. Voir aussil'excellent documentaire d'Arte : Comment Trump a manipulé l'Amérique, octobre 2018 (disponible en ligne jusqu'au 6 janvier 2019).

La société a notamment analysé les profils d'électeurs sur Facebook dans les swing states lors de la présidentielle américaine de novembre 2016. Cela a permis ensuite à d'autres équipes périphériques à la campagne de Trump de cibler des populations de swing states avec des fakes news fabriquées par des sites conspirationnistes « alt-right », sans compter l'effet de relai des bots crées par des équipes financées directement ou indirectement par la Russie poutinienne et mafieuse.

Si on allait plus loin, on pourrait imaginer l'utilisation de réseaux de neurones génératifs fabriquant de fausses vidéos de personnalités, pour influencer l'opinion qu'en ont les électeurs. Pas besoin d'AGI pour y arriver !

Il suffit d'exploiter les technologies existantes et d'avoir de mauvaises intentions. Le mal, c'est l'Homme ! Yuval Harari s'inquiète à juste titre de l'usage de l'IA par des dictatures[855].

- L'exploitation sémantique des analyses terrains réalisées par les volontaires de « En Marche », par **Proxem** (2007, France). Cette société avait produit juste après l'annonce de la candidature d'Emmanuel Macron une longue présentation montrant quels mots clés ressortaient des enquêtes terrain. Cela avait l'air d'exploiter des techniques assez basiques de traitement du langage.
- La tentative de prévision de ce que le Congrès US pourrait voter, par un professeur de l'université de Vanderbilt, J.B. Ruhl et le développeur John Nay dans PredictGov[856]. Le site n'est plus en ligne et n'a probablement pas prévu le vote de John McCain au Sénat américain contre la suppression d'Obamacare fin juillet et fin septembre 2017. Il ne pouvait pas non plus prévoir l'issue du vote de confirmation du juge Brett Kananaugh à la Cour Suprême par ce même Sénat en octobre 2018.

Tout cela reste encore artisanal. Mais ce n'est peut-être que la face visible d'un gros iceberg méconnu. Mais qui régulera les usages de l'IA par les politiques ?

Politiques de l'IA

L'agitation des peurs autour de l'IA en a fait un véritable sujet politique. Le phénomène n'est pas nouveau en soi mais l'est par son ampleur. D'habitude, les pouvoirs publics s'emparent de sujets technologiques avec un retard de phase chronique. Ici, ils sont quasiment en avance de phase, en tout cas, relativement aux menaces de l'IA.

La posture politique suit une gradation relativement classique entre tenants d'une innovation Schumpétérienne libérale qu'il ne faut pas tenter de ralentir et ceux d'un Etat régulateur et protecteur cadrant les usages, l'éthique des affaires et l'économie en général. La première est dominante aux USA tandis que la seconde l'est en Europe et surtout en France où l'étatisme ne faiblit jamais, et résiste fort bien aux alternances politiques.

L'IA est aussi une révolution industrielle et les Etats ont compris qu'il ne fallait pas louper le coche. Des deux côtés de l'Atlantique, les Etats sont des plus prolixes en rapports et plans autour de l'IA.

Barack Obama avait été interviewé par Joi Ito dans Wired en août 2016 et articulait déjà une vision claire des enjeux autour de l'IA[857]. Juste après, son administration a produit deux rapports en fin de mandat, le premier The Administration's Report on the Future of Artificial Intelligence, publié en octobre 2016 après une consultation publique faisait quelques recommandations élémentaires : l'IA devrait servir à améliorer le bien public, les gouvernements devraient l'utiliser, l'IA devrait compléter et non pas remplacer les hommes, et l'usage des véhicules autonomes devrait être régulé.

---

[855] Dans Why Technology Favors Tyranny de Yuval Harari, octobre 2018.

[856] Voir Artificial intelligence tries to make sense out of the mess that is Congress, avril 2017.

[857] Voir Barack Obama, neural nets, self driving cars, and the future of the world, dans Wired, octobre 2016. Sa position était bien documentée et articulée. On se met à fantasmer sur une interview comprenant exactement les mêmes questions, mais posées à Donald Trump. L'A/B testing de POTUS serait saisissant !

S'en suivit Artificial Intelligence, Automation, and the Economy en décembre 2016, qui anticipait des bouleversements sur le marché de l'emploi qui peuvent être absorbés pour peu que les efforts adéquats soient lancés côté formation et qu'une réflexion sur la répartition de la valeur ait lieu.

Un comité juridique du Parlement Européen publiait en février 2017, un court rapport European civil law rules in robotics, plaidant pour la création d'un cadre juridique encadrant l'usage des robots mais en s'opposant à une taxe sur les robots.

La France a suivi de peu avec la publiction par le gouvernement d'un rapport et de l'initiative France IA en mars 2017, produits en deux mois concommitamment avec un rapport des assemblées portant principalement sur les questions sociétales et d'éducation soulevées par l'IA. Peu de temps après l'arrivée d'Emmanuel Macron à l'Elysée, le gouvernement a confié au Député Cédric Villani la mission qui porte son nom pour « *édudier les actions nécessaires pour permettre à la France et à l'Europe d'être à la pointe de l'économie fondée sur l'IA* » et pour identifier les moyens de moderniser l'Etat avec l'IA. Le rapport a été remis fin mars 2018 et le gouvernement a confié en juillet 2018 le soin à Bertrand Pailhès de suivre la mise en œuvre du plan présenté par Emmanuel Macron le 29 mars 2018[858].

Au Royaume-Uni, le Parlement a lancé un appel à consultation sur ces mêmes sujets en juillet 2017 alors que sa Commission Science et Technologie avait déjà publié un premier rapport en octobre 2016, assez succinct avec 44 pages, Robotics and artificial intelligence.

D'autres rapports ont été produits par divers groupes de pression sur les Etats, comme en Australie, où le cabinet de conseil AlphaBeta publiait en août 2017 The automation advantage (44 pages, *ci-contre*) qui dénonce le retard des entreprises australiennes dans l'adoption de l'IA et fait miroiter un potentiel économique de $2,2T (trillions = 1000 milliards), probablement un peu surestimé au regard du PIB actuel du pays qui est de $1,2T même si ce potentiel économique est étalé sur quinze ans.

L'une des inquiétudes partagées par les politiques et certains milliardaires de la tech est le déséquilibre que l'IA pourrait créer dans la fabrique sociale des sociétés, détruisant à petit feu la classe moyenne qui assure sa stabilité[859]. Se posent plein de questions autour de l'IA : formation, développement des startups, organisation de la recherche, adoption de l'IA par les entreprises, régulation et protection de la vie privée.

Rapports et plans

Nous allons creuser ici l'Histoire des rapports sur l'IA en France, étalée entre 2017 et 2018 avec le premier plan France IA (mars 2017), le rapport des assemblées (mars 2017), le rapport de la Mission Villani (mars 2018), le rapport de France Stratégie sur l'IA et le travail (mars 2018), le plan de l'Académie des Technologies (avril 2018) et le plan de la région Ile de France (octobre 2018).

---

[858] Voir Discours du Président de la Répubique AI for Humanity, 29 mars 2018 et son interview dans Wired Emmanuel Macron talks to Wired about France's AI strategy, mars 2018.

[859] Voir How Technology wrecks the middle class de David Autor et David Dorn, aout 2013 et Why governments need to respond to the Fourth Industrial Revolution de Iain Klugman, septembre 2018.

plan France IA du gouvernement
mars 2017

rapport de l'OPECST
Sénat+Assemblée
mars 2017

rapport Mission Villani
Assemblée Nationale
mars 2018

France Stratégie
IA et travail
mars 2018

Région Ile de France
Plan IA 2021
octobre 2018

Plan France IA

Je faisais un tour d'horizon de ce plan et rapport de 350 pages dans un article publié en mars 2017[860] et dont voici un résumé légèrement actualisé.

Ce plan décrivait une recherche fondamentale assez dispersée et centrée sur la recherche publique et d'un manque de transferts technologiques, ce qui n'est pas une spécificité de l'IA. Il évoquait le rôle stratégique des données qui alimentent l'IA. On y trouvait une cartographie de la recherche française en IA complété par celle des entreprises l'utilisant.

Il comprenait des propositions d'orientation de cette recherche, notamment dans l'IA symbolique sous toutes ses formes qui complète l'IA connexionniste qui domine l'univers du deep learning pour la vision et le traitement du langage. Le rapport détaillait aussi les stratégies de grands acteurs américains et chinois ainsi que les plans lancés par différents gouvernements dans le monde.

Le plan proposait de se focaliser sur sept marchés : l'automobile, la relation client, la finance, la santé, les énergies renouvelables, la robotique et l'éducation numérique.

Le plan d'action prévoyait notamment la candidature à un *projet phare de technologie émergente* de l'Union Européenne (« FET flagship ») du type du Human Brain Project, sur l'IA, pouvant être financé à hauteur de 1Md€, le lancement d'un programme IA dans le cadre du Plan pour les Investissements d'Avenir (PIA 3), le financement d'une infrastructure mutualisée de calcul de puissance en IA pour la recherche et la création d'un consortium public-privé sur l'intelligence artificielle (devenu le Hub France IA en septembre 2017).

Le plan oubliait, semble-t-il volontairement, les composants électroniques, qui sont un domaine où les opportunités autour de l'IA sont significatives avec au moins deux technologies clés : les processeurs neuromorphiques et les processeurs quantiques généralistes qui pourraient transformer radicalement le paysage informatique autour de l'IA dans les deux décennies à venir.

Rapport des assemblées

Publié quasiment simultanément au plan France IA en mars 2017, le rapport de l'Office Parlementaire d'Evaluation des Choix Scientifiques et Technologiques[861] « **Pour une intelligence artificielle maîtrisée, utile et démystifiée** » complétait le plan France IA de l'exécutif en abordant surtout la dimension sociétale et réglementaire.

---

[860] Que j'ai commenté ici au moment de sa publication : Les hauts et les bas du plan France Intelligence Artificielle, mars 2017.

[861] Qui associe députés et sénateurs. Le mathématicien Cédric Villani a pris la présidence de cet office depuis son élection comme député LREM en juin 2017.

Nourri par des déplacements à l'étranger (USA, Royaume Uni, Suisse) et de nombreuses rencontres, le rapport démarre avec un très bon panorama de l'histoire de l'intelligence artificielle qui remet pas mal de pendules à l'heure, même s'il comprend quelques perles. On y découvre aussi les 12 laboratoires impliqués dans l'IA au CNRS (GREYC, IRIT, LAAS, LAMSADE, LATTICE, LIF, LIG, LIMSI, LIPN, LIRMM, LORIA, LRI, TIMC) totalisant environ 300 chercheurs permanents.

Le rapport comprend surtout une quinzaine de propositions regroupées en trois parties reprenant les éléments du titre du rapport :

Côté **maîtrise**, on sent poindre de relents de principe de précaution avec des oxymores tels que la volonté d'éviter une régulation de la recherche en intelligence artificielle tout en voulant favoriser une IA sûre, transparente et juste via des chartes, de la formation à l'éthique de l'IA, la création d'assurances spécifiques aux robots et la création d'un Institut National de l'Ethique de l'IA et de la Robotique. Bref, une IA morale comme un capitalisme moral, qui n'existe déjà pas ? La maîtrise se veut aussi sociale avec de la formation continue pour adapter le marché du travail aux mutations de l'IA (on pourrait dans pas mal de cas y caser le numérique en général).

Côté **utile**, il est question de valorisation de la recherche fondamentale, et de la rendre plus transversale, d'encouragement à la création de champions européens de l'IA, d'orientation des investissements vers des applications sociétalement utiles, de création de cursus de formation sur l'IA[862] et aussi de promotion de la diversité et de la place des femmes dans la recherche en IA[863]. Le rapport préconisait la création de projets de recherche transversaux structurants. Le rapport préconisait aussi la création de champions européens. Mais cela ne se décrète malheureusement pas.

Et côté **démystification**, re-formation, mais dans le primaire et le secondaire, de la sensibilisation du grand public, la création d'un salon international de l'IA et de la robotique.

Rapport de la Mission Villani

Le Rapport de la Mission Villani était publié le 28 mars 2018 à l'occasion d'une belle conférence « AI for Humanity » organisée au Collège de France avec un beau panel d'intervenants français et étrangers comme Dennis Hassabis de DeepMind et bien évidemment, Yann LeCun de Facebook.

J'avais abondamment commenté le contenu du rapport dans Ce que révèle le Rapport Villani, en mars 2018[864]. Un œil exercé et un bon sens pratique ne peuvent en général qu'être déçus par ce genre de rapport et celui-ci ne faisait pas exception.

---

[862] Je remarque que les cursus français ne donnent pas toujours lieu, comme dans les principaux cursus anglo-saxons, à la publication des supports de cours en ligne en libre accès. Voilà une belle zone de progrès ! Comme à Stanford pour ce cours de traitement du langage : http://web.stanford.edu/class/cs224n/syllabus.html, le cours de Stanford de reconnaissance d'images d'Andrej Karpathy : http://cs231n.stanford.edu/syllabus.html, ou ce cours généraliste sur l'IA de l'Université d'Amsterdam dont les supports de présentation sont très bien faits : http://uvadlc.github.io. Heureusement, certaines universités font de beaux efforts. Comme l'Université de Lyon 2 qui publie tous les supports de son cours de data mining et data science.

[863] Un écueil qui est aussi soulevé dans le rapport AIReport de l'initiative **ArtificialIntelligenceNow** qui évoque le manque de diversité des chercheurs et développeurs de l'IA. Non seulement, les femmes sont sous-représentée mais aussi les minorités : « *Like all technologies before it, AI systems reflect the values of their creators, and there is hope that increased diversity among those developing, deploying, and maintaining AI systems may help create a future in which these technologies promote equality. Currently, however, women and minorities continue to be under-represented in the field of AI particularly, and in computer science overall* ».

[864] A de rares exceptions près, l'accueil sur le rapport Villani était positif dans la presse française. Elle se félicitait que le sujet soit pris en compte et que l'éthique et les dimensions sociétales y soient mises en avant. Les réactions en creux comme la mienne n'étaient pas nombreuses. On peut citer IA et éthique: le contresens navrant de Cédric Villani de Philippe Silberzahn, Réaction au rapport Villani d'Alain Mueller qui portait sur le manque d'approche entrepreneuriale autour de la notion clé de product management, Intelligence artificielle, être suiveur ou devenir leader de Gilles Babinet, puis Et si l'intelligence artificielle n'était pas le problème du rapport Villani ? de Ariel Kyrou et Thierry Taboy qui posent la question sur la société que nous voulons créer avec l'IA.

Le plan se focalisait surtout sur l'organisation de la recherche en France avec la création d'instituts 3IA voisins dans le principe des IRT, la revalorisation du salaire des chercheurs en IA[865], l'adoption de l'IA par la puissance publique et les questions d'éthique. Il était très tourné sur les besoins du marché français. Le rapport manquait de pédagogie sur la dimension entrepreneuriale, n'insistant pas assez sur les spécificités des startups dans l'IA et les difficultés habituelles des entrepreneurs français à créer des produits.

Etaient tout aussi décevantes les parties sur les composants (introduction sur les chipsets neuromorphiques et proposition de créer un supercalculateur pour les chercheurs), sur la robotique (un plan franco-allemand vide comme l'éther), sur la santé (avec une vision trop franco-française des besoins) ou les véhicules autonomes (pas d'ambition clairement affichée). La partie formation était convenable, tout comme la volonté affichée d'attirer plus de femmes dans les filières de l'IA.

On peut se poser la question de la mesure de notre performance économique dans l'IA. La comparaison avec les USA et la Chine n'a pas de sens au vu de la taille de leurs économies respectives. Quels inicateurs (KPI) indiqueraient que nous sommes bons dans l'IA en tant que pays ? Une production de valeur ajoutée à base d'IA supérieure au poids du PIB de la France dans le PIB mondial ? Un solde commercial positif dans le domaine ? Personne ne s'est encore aventuré à imaginer comment mesurer tout cela ! Cela peut aussi se mesurer en nombre d'unicorns, d'introductions en bourse (IPO), voir idéalement en chiffre d'affaires de startups de l'IA. Mais comme dans tout plan public, le Rapport Villani ne propose aucun indicateur avec un horizon de temps ! Bref, on ne sait pas vraiment qualifier et quantifier les objectifs à atteindre ni donc les résultats.

Rapport de France Stratégie

Commandé par Muriel Pénicaud, ministre du Travail et Mounir Mahjoubi, secrétaire d'État chargé du numérique, le Rapport de France Stratégie sur l'Intellgence et le travail[866] fait une analyse circonstanciée de l'impact de l'IA sur plusieurs secteurs : les transports, la banque et la santé.

Le rapport est très mesuré et évite de sombrer dans le catastrophisme habituel. Pour les rapporteurs, l'impact de l'IA sur l'emploi dans ces différents sera à la fois lent et modéré.

Plan de l'Académie des Technologies

Quelques jours après la publication du rapport de la mission Villani, l'Académie des Technologies remettait le couvert avec son propre plan IA, coordonné par Yves Caseaux, l'un de ses membres et également DSI de Michelin.

Le plan se veut plus pratique et tourné vers les besoins des entreprises. Il souhaite aussi promouvoir la dimension européenne de l'IA tout comme sa dimension éthique[867]. Il fait aussi de la pédagogie en décrivant les différents outils de l'IA destinés aux entreprises.

---

[865] La proposition consiste à doubler le salaire des chercheurs rémunérés par l'Etat débutants sachant qu'ils démarrent au CNRS à environ 36K€ bruts tandis qu'un ingénieur en IA chez Google démarrerait à 140K€. Mais il ne faut pas confondre chercheur et ingénieur. Le décalage entre les deux a toujours été élevé, quelle que soit la discipline. Le hic vient de ce que les GAFA recrutent aussi des chercheurs et au prix des ingénieurs, et cela a un effet inflationniste certain. Le secteur public ne peut pas suivre en Europe. Aux USA, il en va différemment car nombre d'universités qui emploient des chercheurs sont privées et ont plus de souplesse. Mais le salaire des chercheurs aux US n'est pas mirobolant non plus. Il n'a pas l'air de dépasser les $80K ce qui équivaut en intégrant le coût de la santé et de l'éducation à environ 50K€ en France. Dans les chiffres brandis par les uns et les autres sur les effectifs de R&D des grands acteurs du numérique, on confond souvent chercheurs et ingénieurs. Ces derniers sont la majorité dans les GAFAMI (les géants américains) et les NATU (les géants chinois) avec un rapport qui est souvent de 1 à 20 entre le nombre de chercheurs et d'ingénieurs/développeurs.

[866] Voir Intelligence Artificielle et Travail, France Stratégie, mars 2018 (90 pages).

[867] Voir le rapport Renouveau de l'intelligence artificielle et de l'apprentissage automatique et la présentation associée d'Yves Caseau, avril 2018.

Il insiste sur le manque de compétences dans l'ingénierie des projets d'IA et propose la création de laboratoires d'essai et de certification de solutions d'IA par marché vertical tout comme la création d'un observatoire de l'Intelligence artificielle, placé sous le contrôle d'une agence européenne (à créer), pour suivre les bonnes pratiques de l'IA dans la société.

Plan de la région Ile de France

En octobre 2018, la région Ile de France présentait son plan sur l'intelligence artificielle, conçu comme un rebond sur les propositions de la mission Villani. Il ambitionne de faire de la région le hub de l'IA en Europe en s'appuyant notamment sur l'écosystème d'enseignement supérieur et de la recherche publique et privés. Comme dans de nombreux plans, on y trouve à la fois une logique de dynamisation de l'offre notamment des startups et de la demande, en particulier des PME.

Le plan s'articule sur quinze mesures dont la praticité est variable d'un cas à l'autre.

Un premier gros volet vise à dynamiser les usages de l'IA dans les PME. Les mesures ne sont pas évidentes à déployer : des packages de conseil et de PoC (projets pilotes) pour une centaine de PME, la création d'un inventaire de l'offre d'IA, la mutualisation des données industrielles, l'accès à une puissance de calcul souveraine (avec des GPU Nvidia), le projet Inriatech qui veut proposer une offre de recherche en IA à destination ldes PME et ETI, des formations BAC+2 pour les jeunes et les demandeurs d'emploi réalisées par Simplon.co pour Microsoft et la création d'un lycée de l'IA pour l'apprentissage du code dans l'IA (ce qui me semble un peu trop tôt, et pas associé à un Bac spécifique).

S'ensuit un soutient soutien financier au projet Digihall piloté par le CEA consistant à créer un écosystème d'innovation sur le plateau de Saclay, le développement des coopérations internationales avec le Québec, la Bavière et la Corée du Sud et la communication sur l'excellence de l'IA francilienne.

Enfin, et c'est le plus concret et réalisable, des challenges thématiques destinés aux startups dans l'oncologie et l'Hôpital du futur, dans l'industrie autour du transfer learning, les politiques publiques. Les lauréats du premier challenge IA étaient décernés le 15 octobre 2018 avec des prix conséquents étant de 100K€ (Panda, Smartify), 350K€ (LightOn) et 700K€ (Therapanacea).

Ethique

Le Rapport Villani s'est largement fait le relai du besoin de limiter les dérives du far-west de l'IA en mettant en avant la notion d'IA éthique. Ce concept aurait très bien pu être accolé aux logiciels depuis qu'ils existent. L'approche probabiliste du deep learning a mis son grain de sel dans l'équation pour justifier de s'intéresser de plus près aux conditions de la mise en œuvre de l'IA. C'est anodin lorsque le machine learning permet à Netflix de nous recommander une série B correspondant à nos goûts.

Ca l'est moins pour une IA qui gère un diagnostic médical, pilote un véhicule autonome ou décide de nous octroyer un prêt bancaire. La question est encore plus épineuse pour les robots qui interagissent avec les humains, surtout les robots humanoïdes.

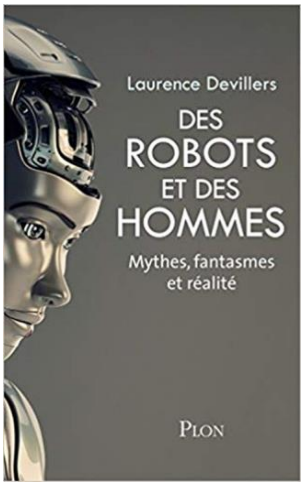
La tendance est donc de faire de l'IA éthique « by design » comme on fait de la « privacy by design », les deux notions étant d'ailleurs étroitement liés. L'IA by design requiert des objectifs de moyens pour les créations d'applications exploitant de l'IA. La question clé pour les gouvernants est de savoir si l'on met en place une règlementation a priori ou a posteriori de la création des innovations.

L'éthique de l'IA couvre en fait plusieurs aspects que les entreprises doivent intégrer dans leurs réflexions, organisatin, création de solutions intégrant de l'IA et sur leur communication :

- Les **règles d'éthique** appliquées (ou pas…) par les IA et les robots dans leurs interactions avec les humains, ce qui intègre aussi l'évitement de biais divers des algorithmes et des données servant à entraîner les IA.
- Le **respect de la vie privée** au sens large du terme.
- La **diversité des équipes** qui créeent ces solutions, notamment dans l'équilibre entre hommes et femmes.
- La **pédagogie** sur ce qu'est et fait l'IA vis-à-vis du grand public.

Règles d'éthique

L'une des spécialistes de cette question en France est **Laurence Devillers**, chercheuse au CNRS et auteur du livre « Des robots et des hommes ». Elle met en avant divers risques à venir : le risque émotionnel lié à un attachement trop fort aux robots[868] ou au contraire un rejet du robot et celui d'une société de plus en plus paresseuse, où l'on cèderait trop souvent son libre arbitre aux machines au point de permettre un contrôle de sa vie pouvant devenir totalitaire.



Cela pourrait créer un gouffre encore plus béant qu'aujourd'hui entre ceux qui connaissent les entrailles des systèmes et la majorité des utilisateurs. La question se poste pour les robots ayant une apparence physique ainsi que pour les logiciels sans incarnation physique autre que du texte, des avatars en 2D ou de la voix. Nombre de ces questions se posent même sans IA, comme dans les usages des mobiles ou dans les jeux vidéo.

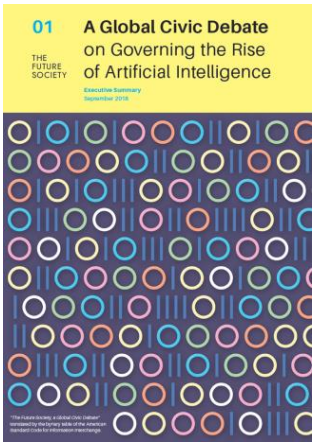
Quand Snapchat créé l'addiction chez les adolescents en réduisant leur popularité sur le réseau social s'ils ne publient rien, il créé un système pavlovien asservissant. Et pourtant, sans IA. Laurence Devillers propose la mise en place de règles de bonne conduite pour que les IA et les robots soient éthiques dès leur conception, pour éduquer les utilisateurs et surtout les plus jeunes, la création d'outils de vérification de ces règles.

Cela irait jusqu'à la création d'un cadre légal et pénalisant en cas de non-respect de ces règles, le tout étant encadré par des comités éthiques indépendants. Sa praticité ne serait pas triviale.

L'un des défis est de concilier les modèles économiques des acteurs du marché avec ces éléments d'éthique, notamment lorsque sont mis en œuvre des modèles publicitaires et addictifs. Même si le respect de l'éthique ne devrait pas être dépendant des modèles économiques de ces acteurs.

[868] Voir Robot : une étude prouve qu'on ressent des sentiments pour une IA, de Pierrick Labbé, août 2018.

Le souci de l'éthique de l'IA n'est pas spécifique à la France. C'est une préoccupation mondiale, ou tout du moins occidentale. Nombre d'associations ou consortiums d'origines américaines ont été créées dans cette mouvance comme **Open AI**[869], **The Future Society**, **Partnerships on AI** (créé par les GAFAMI) ou encore **AINow**.



La déclaration de Montréal de mars 2018 est une initiative lancée par l'**Université de Montréal**. Elle résulte d'un processus de « co-construction » qui faisait suite au Forum on the Socially Responsible Development of Artificial Intelligence, de novembre 2017.

The Future Society[870] publiait en septembre 2018 A global civil debate on governing the rise of artificial intelligence (13 pages), une proposition de gouvernance de l'IA.

Dans le lot, il y a aussi le **Future of Life Institute** créé par Max Tegmark, l'auteur de Life 3.0 qui décrit un futur plein d'IA et les risques associés. Y sont définis les principes d'Asilomar, du nom d'une conférence organisée en janvier 2017 en Californie avec une centaine de participants représentant le gothat occidental de l'IA[871]. Les principes recouvrent : la robustesse des IA, le financement équilibré d'IA et de l'analyse de leur impact, un impact social équilibré de l'IA, des IA auditables (ou explicables), un alignement des IA sur les valeurs humaines, le droit à la vie privée, l'évitement d'une course aux armes autonomes, le développement d'éventuelles AGI devrait être géré avec une infinie précaution surtout pour des versions s'améliorant de manière récursive.

Les détracteurs de ces démarches rappellent que les notions d'éthique sont très variables selon les pays, cultures et religions. En Chine par exemple, l'harmonie d'ensemble de la société est plus importante que les libertés individuelles. D'où le système politique qui y sévit et les libertés limitées, comme dans l'usage sans vergogne de la vidéo-surveillance ou la notation des citoyens dans leurs pratiques sociales.

Il existe une autre notion d'éthique, plus générale, qui consiste à utiliser l'IA pour des solutions qui sortent de la sphère économique traditionnelle et visent le bien social, notamment en faveur des populations les plus défavorisées. C'est la « tech for good » et l'entrepreneuriat social et solidaire (ESS) qui ont sa déclinaison « **AI for good** », correspondant même à une initiative du même nom sous l'égide des Nations Unies[872].

Du côté des entreprises, l'IA va faire rapidement partie des obligations de leur Responsabilité Sociale et Environnementale. Elles devront afficher la manière dont elles intègrent l'éthique de l'IA dans les solutions qu'elles exploitent aussi bien à destination de leurs salariés que de leurs clients.

Vie privée

La notion de vie privée ramène immédiatement à l'application du **RGPD** en Europe, cette règlementation dont, en tant qu'utilisateurs, nous pouvons apprécier au quotidien l'absurdité avec ces validations pavloviennes de cookies qui nous sont imposées par tous les sites web que nous visitons. La perte de temps associée est significative.

---

[869] Voir OpenAI Director Shivon Zilis explains why AI requires oversight now de Darrell Etherington, octobre 2017.

[870] Voir leur site http://www.thefuturesociety.org/. C'est une ONG avec quelque Français dans ses advisors : Brice Lalonde, Hubert Védrine advisor et Nohza Boujemaa de l'INRIA.

[871] Voir The Asilomar AI Principles et la conférence Beneficial AI 2017.

[872] Voir le AI for Good Global Summit organisé par l'ITU à Genève en mai 2018. Une déclinaison française était lancée à l'Hôtel de Lassay sous l'égide de France is AI, France Digitale, et en présence de Cédric Villani le 3 avril 2018. Le principal sponsor de l'événement était … Facebook, en pleine affaire Cambridge Analytica. Des associations liées à la tech for good étaient aussi de la partie : la Croix-Rouge française, Bibliothèques sans frontières, Green Cross tout comme la startup DataforGood qui est notamment associée à la startup Bayes Impact de Paul Duan.

Derrière cet agacement se cachent des procédures que les entreprises doivent adopter pour préserver la confidentialité et la sécurité les données collectées sur les utilisateurs. Ce sont des objectifs de moyens clairement documentés.

Qu'en est-il de ces données une fois qu'elles passent à la moulinette de l'IA ? Lorsqu'elles sont exploitées par des briques de machine learning et de deep learning, elles perdent généralement leur caractère nominatif. Il n'y a pas plus anonyme qu'un gros réseau de neurones entraîné avec des paramètres décrivant le comportement d'un grand nombre de clients. Ces réseaux de neurones servent ensuite à faire des classifications et des prévisions sur le comportement de nouveaux clients ou de clients existants, basés sur le comportement collectif.

La préservation de la vie privée n'est pas garantie avec des données anonymisées dès lors qu'elles intègrent plus d'une dizaine de paramètres[873]. Notre signature personnelle est en effet unique dès que l'on combine une masse critique de ces paramètres. Si certains de ces paramètres sont communs avec ceux d'une autre base, elle non anonymisée, alors, on peut faire des recoupements et désanonymiser une base de donnée anonymisée[874].

Le principe de la « *differencial privacy* » doit permettre d'éviter qu'une entrée de donnée anonymisée d'un utilisateur ait un impact significatif sur les résultats de l'algorithme, permettant son identification éventuelle en croisant le résultat avec une autre base de donnée. Par exemple, si on cherchait à obtenir une statistique sur un échantillon d'utilisateurs qui est trop réduit, on pourrait obtenir l'information associée sur un utilisateur particulier de cet échantillon. Les techniques couramment utilisées consistent à ajouter du bruit dans les données anonymisées qui alimentent le modèle. Du bruit est surtout ajouté en sortie pour éviter que le modèle puisse être utilisé pour identifier des utilisateurs particuliers. C'est particulièrement important pour les applications dans la santé ou les assurances[875]. Apple et Microsoft sont de grands promoteurs de la *differencial privacy*.

On peut aussi alimenter des modèles de machine learning avec des données chiffrées. Les méthodes de chiffrement dites homomorphes permettent de le faire en théorie. Un tel chiffrement permet de réaliser des opérations mathématiques comme des additions et multiplications sur des données chiffrées et de déchiffrer ensuite le résultat.

Cela permet d'alimenter des algorithmes avec des données chiffrées, ce qui peut-être utile pour des services réalisés dans le cloud. Une entreprise ou organisation peut les entraîner les modèles avec des données chiffrées et exécuter les modèles entrainés avec des entrées et des sorties chiffrées. Les opérations de chiffrement et déchiffrement ont lieu chez le client. L'opérateur du cloud ne peut ainsi pas accéder aux données et exploiter le modèle entrainé.

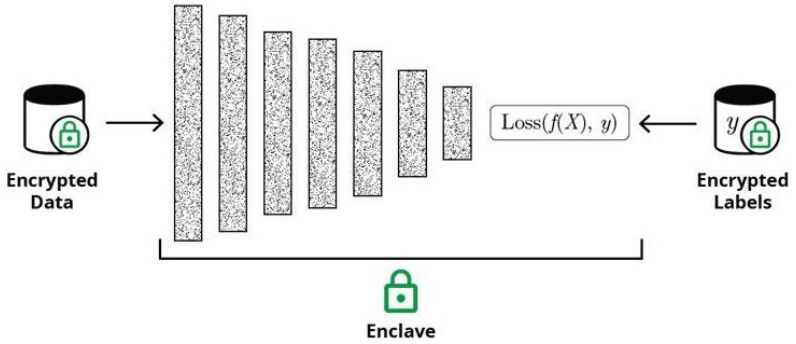
---

[873] Voir Artificial intelligence and privacy Report, du norvégien Datatilsynet, janvier 2018 (30 pages) qui fournit une approche européenne sur la manière de tenir compte du RGPD dans sa stratégie de données et d'IA ainsi que Artificial intelligence and privacy Issues paper, Office of the Victorian Commissioner, juin 2018 (15 pages).

[874] La désanonymisation des données privées et les risques associés sont les thèmes de recherche de Sonia Ben Mokhtar, du CNRS LIRIS à l'Insa de Lyon. Voir son site https://sites.google.com/site/soniabm. Dans Artificial Intelligence as as Digital Privacy Protector, 2017 (19 pages) d'Andrea Scripa Els de Harvard explique comment l'IA peut améliorer la vie privée et contourner ces problèmes. Avec notamment la notion de « Differencial Privacy ». Le processus de réidentification d'utilisateurs est bien décrit dans la présentation Big Data, Artificial Intelligence and Privacy de Stephen Kai-yi Wong (responsable de protection de la vie privée à Hong Kong), 2017 (34 slides). Dans la même lignée, Latanya Sweeney est connue pour ses travaux sur la collecte d'informations privées par les applications mobiles. Voir Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps, octobre 2015. Elle est aussi la fondatrice du Privacy Data Lab à Harvard.

[875] La méthode a été créée par Cynthia Dwork en 2006. Voir Differential Privacy A Primer for a Non-technical Audience, février 2018 (41 pages), et la présentation associée, 2017 (45 slides), Deep Learning with Differential Privacy, 2016 (14 pages) et Differential privacy and machine learning (127 slides).

C'est un domaine en plein devenir. Les premiers systèmes entièrement homomorphes sont apparus en 2009[876]. Seuls quelques types de machine learning et deep learning peuvent être alimentés avec des données chiffrées de manière homomorphe, comme les Crypto-Nets.



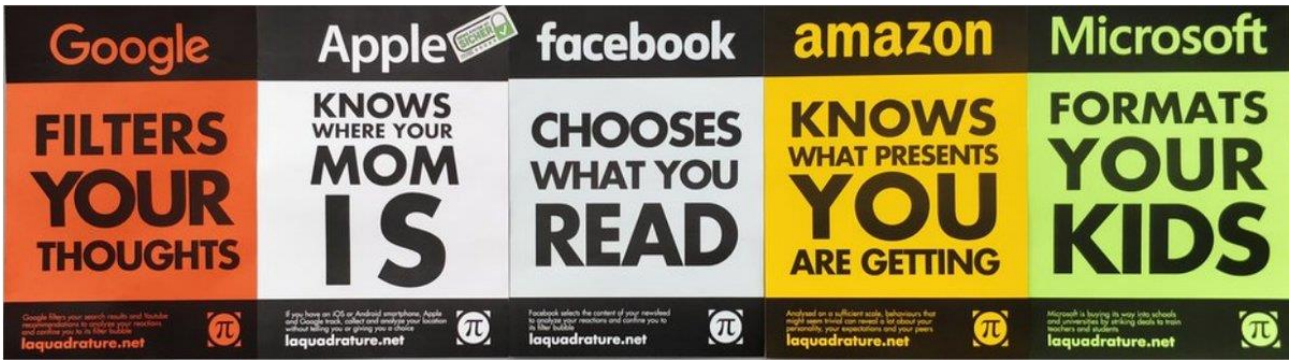
Le véritable risque, associé au fameux biais des données d'entraînement, est que votre cas particulier ne soit pas représenté en quantité suffisante d'un point de vue probabiliste dans ces données d'entraînement ou qu'au contraire il soit surreprésenté.

On peut cependant aussi entraîner des IA avec uniquement vos données personnelles si elles sont en quantité suffisante, par exemple, pour vous faire des recommandations de produits, contenus et services. C'est aussi utilisé dans certaines applications de formations en ligne personnalisées.

Les données des utilisateurs sont généralement anonymisées avant d'entraîner des IA probabilistes. Ces bases intermédiaires ne sont pas directement utilisées. Ce sont des sous produits d'un processus au bout duquel on exploite des modèles d'IA probabilistes entraînés. Les données amont de l'IA sont donc plus sensibles que les modèles entraînés avec.

Comme le rappelle ce visuel issu de la **Quadrature du Net**, les grands acteurs de l'Internet savent beaucoup de choses sur la vie des utilisateurs. Heureusement, à moins d'être peint entièrement en Google ou en Apple, nos données privées sont généralement réparties dans plusieurs de ces systèmes et ils ne se parlent pas (encore). Mais qu'en serait-il dans une ville intelligente très intégrée[877] ?

Une question clé subsiste : comment faire d'un respect scrupuleux de l'éthique et de la vie un avantage marketing capable de renverser la situation ? Dans des marchés existants, cela semble difficile tellement les inerties sont grandes pour faire changer les habitudes tant des consommateurs que des entreprises, surtout si elles sont interdépendantes d'un réseau et d'un écosystème dense de solutions tierces parties. C'est ce qui explique la stabilité de l'empire Microsoft dans les serveurs et les postes de travail depuis plus de 35 ans, celle d'Apple dans les smartphones depuis 11 ans, celle de Google dans les moteurs de recherche, depuis maintenant 20 ans et de Facebook dans les réseaux sociaux.



[876] Voir Crypto-Nets: Neural Networks over Encrypted Data, 2015 (9 pages), la présentation An Introduction to Homomorphic Encryption for Statistics and Machine Learning de Louis Aster, 2015 (68 slides) et Deep Learning on Private Data de Sadegh Riazi, Bita Darvish Rouhani et Farinaz Koushanfar (9 pages).

[877] Voir Utopian Vision, Dystopian Reality, 2017 (25 pages) de Privacy International qui s'inquiète à juste titre des libertés individuelle dans une ville hyper-connectée.

Des acteurs comme le moteur de recherche du Français **Qwant** (2011, France) mettent en avant le respect de la vie privée dans leurs bénéfices clés. Ils ne conservent en effet aucune donnée des utilisateurs pour rendre leurs recherches pertinentes. C'est particulièrement intéressant dans la version de leur moteur de recherche qui est destinée aux enfants.

C'est un avantage concurrentiel en Europe où le souci de la vie privée est le plus élevé, particulièrement en Allemagne et pour des raisons historiques liées (Stasi, Nazis). Certains utilisateurs et organisations – surtout dans le secteur public - abandonnent explicitement Google Search pour cette raison.

Le dilemme est que cela ne permet pas de cibler les utilisateurs de manière aussi pertinente qu'en conservant leurs données. D'où un revenu par utilisateur qui est structurellement plus faible. Cela rend difficile une concurrence face à Google d'un point de vue macro-économique.

Non seulement, ils bénéficient d'économies d'échelle plus grandes mais également d'un bien plus gros revenu par utilisateur (ARPU), nonobstant le fait que l'ARPU de Qwant n'est pas publié. Les solutions pour contourner cela ? Faire de ce service un « utility » régulé voir financé par les pouvoirs publics et veiller à garantir la concurrence de ce marché qui ne devrait pas tolérer qu'un seul acteur d'arroge plus de 90% du marché en Europe.

Il est diffice de changer la donne d'un marché établi avec uniquement cette fonctionnalité de la privacy. Le coût de l'éducation du marché est très élevé[878]. Les inerties d'écosystèmes pèsent tout autant. Il vaut mieux être les premiers à générer des ruptures technologiques ou d'usages pour transformer radicalement un marché. Donc, pour de nouveaux usages, l'approche éthique/vie privée pourrait être plus porteuse, comme dans la santé, la ville intelligente ou les transports.

Il n'empêche que l'adoption de pratiques éthiques dans le numérique, avec ou sans IA, est incontournable pour les startups comme pour les entreprises établies. C'est un impératif sociétal qu'il sera de plus en plus difficile de contourner, soit du fait de la régulation, soit du fait de la réaction de la société en général comme l'a montré l'affaire Cambridge Analytica vis-à-vis de Facebook.

Diversité

L'une des questions clés de l'éthique est l'évitement du biais des algorithmes et des données dans l'entraînement des IA probabilistes, à base de machine learning et de deep learning. L'un des moyens d'y parvenir consiste à rendre les équipes qui la créent plus diverses, à commencer par l'équilibre des genres qui est très mauvais à peu près partout dans le monde.

Que cela repose ou pas sur des stéréotypes, les hommes et les femmes ont en moyenne une approche différente des besoins applicatifs. Les solutions créées principalement par des hommes présentent des biais, ne serait-ce que thématiques[879]. Elles ne peuvent pas prétendre incarner l'ensemble des besoins de la société. C'est l'une des raisons pour lesquelles il faut plus de diversité dans les équipes de chercheurs, d'ingénieurs et de cofondateurs de startups.

Diverses initiatives visent à encourager plus de jeunes filles à s'orienter vers les métiers scientifiques et en particulier ceux du numérique, comme la **Fondation Femmes du Numérique**, lancée fin juin 2018 à la Cité des Sciences et de l'Industrie de la Villette en présence de quatre ministres. Elle servira à financer les actions sur le terrain de plus d'une quarantaine d'associations dont par exemple **Talents du Numérique**, ex Pascaline. Et il y a fort à faire car on assiste même à une régression avec de moins en moins de femmes qui sont attirées par ces carrières scientifiques et techniques.

---

[878] Voir Intelligence artificielle : l'éthique est-elle soluble dans l'économie ? de Christophe Auffray, décembre 2017.

[879] Nous en avons un cas extrême au Japon avec **Gatebox** (Japon) et sa copine virtuelle, qui permet de rompre la solitude du jeune homme célibataire. Cela s'explique par la complexité des relations sociales intersexes chez les jeunes adultes au Japon (vidéo).

Aucune baguette magique ne va résoudre ce problème instantanément. C'est l'affaire de presque une génération. Cela passe par la modification du rôle des filles et des femmes dans les médias, dans l'éducation, dans les jouets, par la valorisation des domaines d'applications du numérique et de l'IA dans des secteurs tels que la santé et l'environnement, par la valorisation de l'utilité sociale des solutions tout comme par celle de role models féminins inspirants[880]. On peut aussi saluer les initiatives **Women in Machine Learning**[881] (USA) ainsi que **Women in AI**, qui a une branche en France.

Pédagogie

Un dernier point mériterait d'être traité côté éthique : la pédagogie. On ne peut avoir de débats sains sur l'éthique de l'IA que si on comprend les réalités technologiques et d'usages de l'IA.

L'IA est très diverse à ces deux niveaux comme nous avons pu le découvrir dans cet ebook. Or cela manque vraiment. L'IA est fantasmagorique. Sa perception et sa compréhension sont trop l'affaire de futurologues et singularistes qui confondent la réalité scientifique d'aujourd'hui et de demain avec les fantasmes de la science fiction. Trois audiences sont clés pour faire avancer les débats : les politiques, les entreprises et le grand public. Ils ont besoins d'un tronc commun de connaissances et de messages différents. Cela explique notamment quelques initiatives qui veulent faire rentrer le numérique, le code et l'IA dans l'enseignement secondaire. Les approches mêlant institutions et société civiles peuvent marquer les esprits comme cette initiative du collectif et cluster **NaonedIA** à Nantes en juin 2018 (*ci-contre*) qui publiait le Manifeste de Nantes, reprenant la déclaration de Montréal[882].

SOCIÉTÉ. À Nantes, un collectif défend des valeurs dans le domaine de l'intelligence artificielle

L'éthique des algorithmes

Comment traiter les masses de données numériques ? Des Nantais veulent poser des garde-fous éthiques.

L'intelligence artificielle sera un raz-de-marée, qui va toucher tous les domaines, avec des enjeux sociétaux très forts ». Francky Trichet (SE) est adjoint au maire de Nantes chargé du numérique. Il est aussi chercheur en intelligence artificielle à l'université. « *Aujourd'hui, les données sont partout et cela peut créer des situations à risque* », estime-t-il.

Qui contrôle les algorithmes qui traitent nos données ?

Qu'est-ce que « l'intelligence artificielle » ? De nouveaux algorithmes capables d'exploiter les masses de données que nous générons. Santé, déplacements, modes de consommation… Tout peut être passé à la moulinette. « *Les algorithmes sont de plus en plus précis, fins, performants. Ils sont capables de faire de la prédiction, dans le domaine de la santé notamment* », explique Francky Trichet. Mais qui contrôle ces algorithmes ? Pour quels intérêts sont-ils conçus ? Quel regard a-t-on sur l'exploitation de nos données ? À Nantes, une cinquantaine d'acteurs du numérique, chercheurs, startups, grands groupes industriels, ont décidé de constituer un collectif : « Naonedia ». Ensemble, ils ont défini une charte pour défendre des valeurs éthiques dans le développement de l'intelligence artificielle. « *Un autre collectif s'est créé à Lyon* », précise Francky Trichet.

Des Nantais ont constitué un collectif pour promouvoir une intelligence artificielle éthique. Photo PO-D. Lanhein

Les Nantais posent huit grands principes dans un manifeste, notamment le respect de la vie privée. « *Nous exigeons des garanties dès la conception des outils d'intelligence artificielle sur la protection de la vie privée et sur le droit des usagers à ne pas subir contre leur gré l'application dérivée de l'intelligence artificielle* », écrit le collectif Naonedia.

Constituer un pôle de ressources à Nantes

Pourquoi le tissu numérique nantais se saisit-il de ce sujet ? « *Nous avons depuis longtemps des chercheurs et laboratoires qui travaillent sur l'intelligence artificielle*, répond Francky Trichet. *Nous avons aussi des startups émergentes, spécialisées dans ce domaine, comme Owkin qui développe des algorithmes pour la santé. Nous avons ainsi une trentaine d'acteurs dans l'écosystème particulièrement impliqués dans le collectif.* »

Au-delà de la création d'une charte pour poser des garde-fous, Naonedia ambitionne de constituer un pôle de ressources sur l'intelligence artificielle, en mutualisant les compétences locales. Le collectif compte sensibiliser, expérimenter, développer des formations dans les entreprises… Et essaimer dans d'autres métropoles pour déployer « *une intelligence artificielle éthique et populaire, respectant les données privées* ».

Emmanuel Vautier

REPÈRES

Le collectif « Naonedia » réunit une cinquantaine d'acteurs : l'université de Nantes, le CHU ou encore l'Inserm mais aussi des startups comme Akeneo, des groupes industriels tel Airbus, des banques, etc.
Site web : www.naonedia.fr

L'influence médiatique en matière d'IA comme en politique est dominée par des personnalités extrêmes. Nous avons en premier lieu le trouble fête, le poil à gratter et l'épouvantail en chef de l'IA qu'est **Laurent Alexandre**.

Puis des évangélistes tels que **Stéphane Mallard** et enfin, des personnalités modérées et écoutées comme **Yann LeCun**, **Luc Julia**[883] et plus récemment **Alexandre Templier**[884]. Nous avons besoin d'un équilibre entre ces différents points de vue.

---

[880] Voir par exemple Quand l'IA se code au féminin, février 2018 qui fait un inventaire de role models féminins qui ont joué un rôle dans l'histoire du logiciel, 100 women in AI ethics de Mia Dand en octobre 2018 et bien entendu l'initiative Quelques Femmes du Numériques !, projet photo lancé en 2012 et devenu en 2016 une véritable association qui contribue à la présentation de jeunes role models dans les lycées et établissements d'enseignemetn supérieurs (depuis 2017 dans les Pays de la Loire et bientôt en Rhône-Alpes).

[881] Au sujet de Women in Machine Learning, voir The Women Changing The Face Of AI, août 2016.

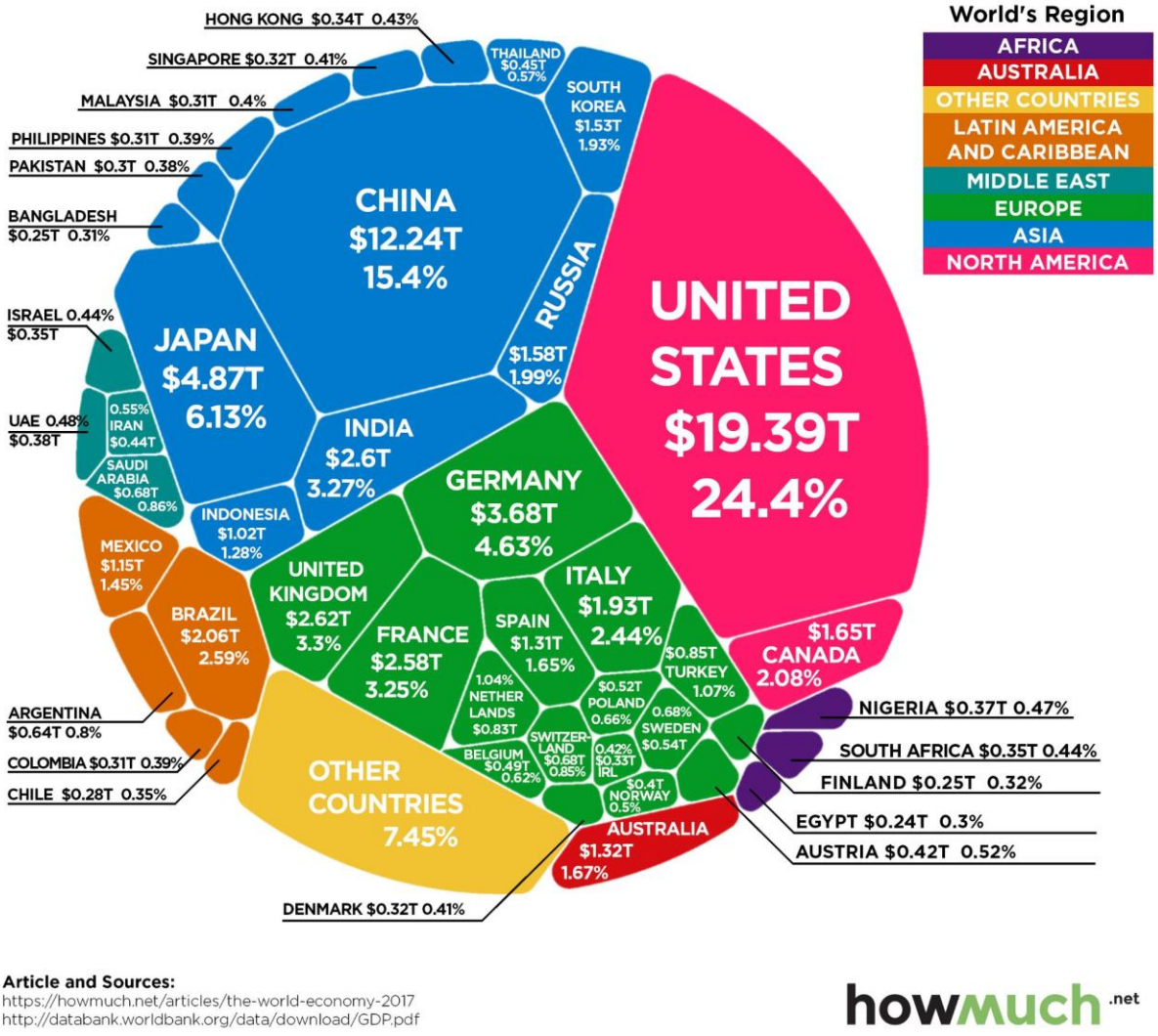
[882] Voir Focus sur NaonedIA, le collectif nantais mettant en avant l'intelligence artificielle pour tous, de Johanna Diaz, juin 2018.

[883] Avec cet ebook, je cible plutôt les entreprises, sans aucune prétention à toucher le grand public.

[884] « L'intelligence artificielle est exposée au risque des technoprophètes », dans Le Monde, octobre 2018.

Nous devons faire preuve de pédagogie sur les réalités technologiques et des usages de l'IA. Un déchiffrage est nécessaire. Il faut ouvrir un peu la boîte noire. Tout comme le moteur à explosion n'explose pas au sens littéral du terme, l'IA est moins dangereuse que ses Cassandre le laissent croire. Aidons le public à décrypter la propagande d'études de cas enjolivées que nous avons pu évoquer dans ce document. Dénonçons l'escroquerie des robots soi-disant autonomes de Boston Dynamics qui ne le sont pas (encore) dans la réalité !

Evitons les généralisations à outrance comme l'interprétation des images médicales par le deep learning qui est extrapolée à tous les actes de santé. N'oublions pas la composante humaine de tous les métiers voués à disparaitre selon les prévisions les plus sombres.



Laurent Alexandre secoue les neurones du monde politique en exagérant le trait, en crédibilisant les thèses de singularistes ou de la connexion hommes machine qu'il sait improbables[885].

Cela fait réfléchir sur de graves problèmes de société qui sont bien réels, et n'ont d'ailleurs pas attendu l'IA pour apparaître. Sur la manière de gérer les inégalités. Sur le rôle de l'école dans la préparation des talents de demain. Sur la fracture des intelligences qui risque de s'agrandir après celle de la fracture numérique.

[885] Avec une bonne moitié de prévisions qu'il fait siennes dans Laurent Alexandre : sept tendances qui annoncent l'intelligence artificielle de 2040, octobre 2018.

Prenons du recul également sur ces peurs exacerbées sur la force des USA et de la Chine dans l'IA alors qu'elle est déjà un fait accompli dans les technologies clés du numérique d'aujourd'hui. Les comparaisons sont toujours abusives côté ordres de grandeur entre un pays de 68 millions d'habitants (la France), de 325 millions (USA) et de 1,4 milliards (la Chine).

Il faut juste se rappeller que nous sommes une puissance moyenne par rapport aux géants chinois et américains. L'Europe est un géant à leur hauteur mais fragmentée. L'approche gagnante ne peut être que transnationale, européenne et même latérale avec d'autres continents ou pays. Mais cette défragmentation est un casse-tête pour l'Europe.

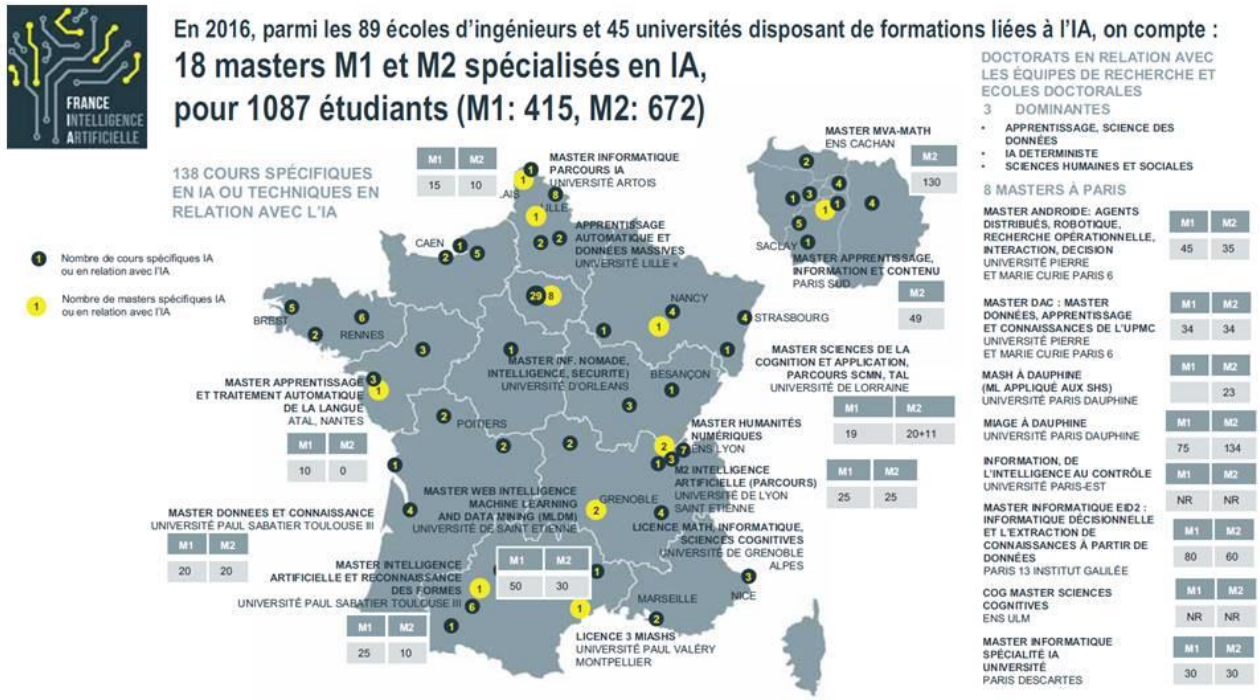
L'Histoire nous a construits fragmentés. Les règlementations européennes et l'Euro n'ont pas suffi à unifier les marchés. Au-delà de questions sociétales et financières, c'est la principale raison de la quasi-inexistance de grands acteurs mondiaux du numérique en Europe (hormis SAP et Dassault Systèmes).

Education

Le Rapport France IA de 2017 faisait un inventaire des formations dans l'IA. Les Masters Spécialisés ne produisaient que 1087 étudiants par an, à comparer avec environ 30 000 développeurs formés par an et qui ne seraient déjà pas suffisants pour répondre à la demande.

La situation va être rapidement tendue sur ce marché. Les progrès de l'IA pourraient être ralentis dans certains pays par le manque de compétences. On ne paramètre par un réseau de neurones convolutionnel ou récurrent en claquant des doigts.

D'où les propositions de ce plan et des plans suivants visant à augmenter le nombre d'étudiants formés à l'IA dans l'enseignement supérieur dans les divers plans gouvernementaux, qui insistent à juste titre sur le besoin de croiser ces formations avec des cursus non informatiques (santé, transports, etc)[886].



[886] L'un des exemples est l'annonce de la création d'un collège (BAC+2) au MIT pour former des jeunes à l'IA en la croisant avec d'autres disciplines et notamment la santé, les sciences politiques et la linguistique. Le collège sera financé à hauteur de $1B, probablement structuré sous la forme d'un endowment dont les intérêts feront tourner le collège sur plusieurs décennies. Il aura 50 enseignants. Le principal donnateur est le CEO du fonds d'investissement Blackstone, Stephen Schwarzman, à hauteur de $350M. Voir Gift of $350 million establishes the MIT Stephen A. Schwarzman College, dans MIT News, octobre 2018 et M.I.T. Plans College for Artificial Intelligence, Backed by $1 Billion de Steve Lohr, octobre 2018.

Nombre d'initiatives de formations à l'IA sont venues du secteur privé [887] sans compter l'autoformation avec des ressources en ligne, particulièrement appréciées par les plus jeunes[888].

La question de l'éducation porte aussi sur toutes les autres filières d'enseignement professionnelles. Comment les rendre plus résilientes à la transformation ou à l'automatisation complète des métiers par l'IA ? C'est ce que traite Laurent Alexandre dans son ouvrage de 2017 « La guerre des intelligences » où il prône un enseignement non spécialisé favorisant la créativité.

Les formations professionnelles (CAP et BAC Pro) ou supérieures forment des spécialistes plus ou moins spécialisés. Ces spécialités n'ont pas empêché des tas de BAC+n de se réorienter dans divers chemins lors de leur vie professionnelle.

Les entreprises ont à la fois besoin de spécialistes prêts à l'emploi et de salariés qui s'adaptent rapidement au changement. C'est une attitude schizophrénique difficile à gérer.

Il faut donc des spécialistes-généralistes... ! C'est-à-dire, des formations suffisamment généralistes et avec une spécialisation permettant quand même de démarrer quelque part. Nombre d'écoles d'ingénieurs et de commerce ont ainsi un tronc commun généraliste et une année de spécialisation. C'est un bon compromis qui mériterait d'être appliqué aux formations professionnelles.

La recherche de talents dans l'IA par les entreprises françaises est exacerbée par la fuite des talents vers les grandes entreprises étrangères qui proposent parfois des salaires mirobolants. Ceux-ci sont cependant souvent exaégérés par l'extrapolation des cas les plus extrêmes[889]. Des salaires compris entre $300K et $500K sont souvent évoqués, en oubliant qu'une part relève de stock options. Ceci étant, les recrutements des entreprises de technologie créent un phénomène de vases communicant qui peut perturber l'activité des chercheurs, en ponctionnant leur capital humain trop brutalement. C'est vrai en France comme aux USA ou ailleurs[890].

Recherche

Dans **The Entrepreneurial State** (2013), Mariana Manzzacuto se bat contre l'idée selon laquelle le secteur privé prend des risques et l'Etat est conservateur et lent. Elle montre qu'au contraire, l'Etat – notamment américain – prend bien plus de risques et investit plus sur le long terme que toute entreprise privée. C'est particulèrement vrai avec l'IA et encore plus vrai dans le cas de l'informatique quantique en France où pour l'instant, une majorité des risques sont pris dans la recherche publique, à l'exception rare d'Atos.

La recherche est à l'origine des grands progrès techniques dans l'IA, matérialisées ensuite par les offres des entreprises de toutes tailles. Geoff Hinton chez Google tout comme Yann LeCun sont d'anciens chercheurs du secteur public au Canada et aux USA. Idem pour les fondateurs de DeepMind et de nombre de startups pointues dans le domaine.

De nombreux frameworks open source ont été créés par des chercheurs du secteur public comme **scikit-learn** qui a bénéficié et bénéficie encore de contributions de l'INRIA.

Au-delà de la recherche fondamentale, on a aussi besoin de plus de recherche appliquée dans tous les domaines et pas seulement dans le numérique « horizontal », ce qui explique par exemple l'implication de l'INRIA dans la santé.

---

[887] Avec notamment le programme de formation proposé par Microsoft et Simplon.co à une cinquantaine de jeunes et de demandeurs d'emploi par an. Mais ce sont des formations BAC+2 assez légères. Elles rendent ces personnes employables dans les activités de services mais pas forcément au niveau attendu par les startups les plus exigeantes qui peuvent de leur côté avoir besoin d'au minimum des BAC+5 quand ce n'est pas des doctorants.

[888] Comme les cours d'IA de Cornell ou de Stanford dont on peut facilement récupérer les supports en ligne avec par exemple Foundations of AI et Machine Learning.

[889] Voir par exemple AI Geniuses Are Being Paid Over $1 Million At Elon Musk's OpenAI de Sam Shead, avril 2018.

[890] Voir Tech firms try to address the risks the AI race poses for research, novembre 2018.

Le plan proposé par la mission Villani consistait à créer des 3IA (Instituts Interdisciplinaires d'IA), des instituts de recherche en IA pluridisciplinaires voisins du concept des IRT (Instituts de Recherche Technologiques) créés à partir de 2009 et à même de faire le pont entre la recherche fondamentale, la recherche appliquée et l'industrie. Il devrait y en avoir moins d'une demi-douzaine qui doivent être choisis à l'issue d'un processus comprend d'abord un appel à manifestation d'intérêt puis un appel à projet, lancés par l'ANR.

Les candidatures étaient celles de Sophia Antipolis (CNRS, Inria et l'Université Nice Côte d'Azur, focalisé sur la santé et le développement des territoires) Toulouse (Université Fédérale de Toulouse Midi-Pyrénées, sur les transports, l'environnement et la santé), Paris (« PRAIRIE », sur la santé, les transports et l'environnement) et Grenoble (« MIAI@Grenoble-Alpes », focalisé sur la santé, l'environnement et l'énergie).

Elles ont toutes été présélectionnées par un jury entièrement international missionné par le gouvernement début novembre 2018. Le processus doit se poursuivre avec une remise de dossier le 15 janvier 2019. Ces 3IA seront financé au maximum à hauteur de 100M€ par an via le PIA3 (Programmes d'Investissements d'Avenir).

Economie et social

Les gouvernements font face à une « perfect storm » en perspective : une révolution technologique en accéléré qui pourrait rapidement bouleverser l'équilibre bancal actuel du marché de l'emploi. Même avec les prévisions les plus optimistes qui anticipent un déficit net de 6% d'emplois d'ici une dizaine à une quinzaine d'années, cela donnerait du grain à moudre.

Supporter une augmentation moyenne de 50% du chômage ne serait pas facile à absorber, même si les prévisions sont très élastiques au niveau de leur échéance et si des pays comme la Grèce et l'Espagne sont déjà passés par là pour d'autres raisons après leurs crises de la dette. Les syndicats commencent à s'en inquiéter, observant l'impact des gains de productivité liés à l'IA sur l'emploi ou sur les objectifs données aux salariés.[891]

Quelles politiques économiques adopter ? Elles tournent toujours autour du développement économique avec les moyens à disposition des Etats pour accompagner les entreprises, en général par le financement de l'amont de l'innovation et de la recherche, et aussi en développant le tissus économique des startups.

L'un des points clés est d'être exportateur de technologies plutôt que simple consommateur. Si la valeur ajoutée de l'IA et les robots viennent d'un nombre réduit de pays, les autres seront toujours désavantagés comme nous le sommes déjà aujourd'hui dans de nombreux pans du numérique, surtout grand public.

Dans **Economic Report or The President**[892], le rapport annuel 2016 sur l'économie de la Maison Blanche publié en janvier 2017 à la fin de l'administration Obama, on découvrait qu'aux USA et en 2013, les startups avaient créé 2 millions d'emplois et les entreprises traditionnelles 8 millions. Donc 20% ! Une proportion énorme sachant que dans le même temps, l'économie française a plutôt détruit des emplois et les startups n'en ont probablement créé que quelques dizaines de milliers tout au plus. Et surtout : la moitié de la R&D fédérale US est dédiée à la défense ! Au milieu des années Reagan (1985), elle en représentait les deux tiers !

---

[891] La CGT organisait un colloque sur l'IA en novembre 2018. J'y intervenais sur le sujet des stratégies industrielles de l'IA, à la suite de Bertrand Pailhes, coordinateur de la stratégie IA de l'Etat, qui présentait la feuille de route du gouvernement. Voir L'intelligence artificielle est l'affaire de toutes et tous de Marie-José Kotlicki, secrétaire de la Ugict CGT, la branche de la CGT pour les ingénieurs, cadres et techniciens. Elle fait un compte-rendu des débats de ce colloque. On peut compléter cela avec le compte-rendu de Christophe Alix dans Libération dans Intelligence artificielle : pour la CGT, le binaire de la guerre suite à mon intervention dans le Congrès de la CGT sur l'intelligence artificielle le 6 novembre 2018.

[892] Voir Economic Report to the President, January 2017 (599 pages).

Cela explique pourquoi tant de projets autour de l'IA sont financés par la l'agence de l'innovation du Pentagone, la DARPA[893]. Y compris trois défis lancés en 2004, 2005 et 2007 sur la conduite automatique, qui ont dynamisé les équipes de recherche de nombreuses universités sur le sujet. Nombre de ces équipes ont été ensuite recrutées par Google pour ses différents projets de voitures automatiques.

En avril 2017, le Secrétaire du Trésor de Donald Trump, Steve Mnuchin, ancien de Golman Sachs, affichait un optimisme étonnant[894], affirmant dans un débat qu'il ne voyait pas de menace sur l'emploi causée par l'IA avant 50 ou 100 ans. Cela lui a valu le sobriquet d'*AI denier*, comme un *climate change denier*. On est passé d'un souci mesuré à une insouciance coupable, mais à la hauteur des compétences de l'actuelle administration américaine. Et cela n'a pas changé depuis un an et demi. Même si on peut être optimiste, il est probable que Steve Mnuchin n'avait pas vraiment réfléchi à la question !

Joi Ito, Scott Dadich, and President Barack Obama photographed in the Roosevelt Room of the White House on August 24, 2016.

Barack Obama, August 2016

*"Low-wage, low-skill individuals become more and more redundant, and their jobs may not be replaced, but wages are suppressed. And if we are going to successfully manage this transition, we are going to have to have a societal conversation about how we manage this. How are we **training** and ensuring the economy is inclusive if, in fact, we are producing more than ever, but more and more of it is going to a small group at the top?"*, Wired, conversation with Jon Ito

Steve Mnuchin, April 2017

"I think it is so far in the future. In terms of artificial intelligence taking over American jobs, I think we're like so far away from that (it's) not even in my radar screen. 50 or 100 more years", Axios News Shapers

En France, le débat politique autour de l'IA a connu un tournant « social » pendant la présidentielle 2017. Il a contribué à la mise en avant de propositions de revenu minimum ou de base[895]. Benoit Hamon et Jean-Luc Mélenchon le justifiaient avec les risques de robotisation des métiers. Une charrue mise avant avant les boeufs, même si les politiques sont parfois des bœufs et sont bien devant.

L'élection de Donald Trump par les fameux blancs de la middle class de la rust belt doit aussi à la désindustrialisation ce ces Etats et à un décalage mal vécu entre ces Etats qui s'appauvrissent et la Silicon Valley qui s'enrichit sans discontinuer.

---

[893] En France, cette mission était historiquement dévolue à la DGA. Mais le Ministère des Armées a créé en septembre 2018 sa propre agence de l'innovation de la défense, dirigée par Emmanuel Chiva et visiblement rattachée à la DGA. Voir Création de l'Agence de l'innovation de défense et nomination d'Emmanuel Chiva au poste de directeur, septembre 2018.

[894] Consulter le transcript exact de son intervention : Read the full transcript from Treasury Secretary Steve Mnuchin's interview, mai 2017.

[895] Voir Today's Artificial Intelligence Does Not Justify Basic Income de Vincent Conitzer, octobre 2016, et c'est toujours valable deux ans plus tard et Les secrets bien gardés du revenu universel de Diana Filipova, août2016, qui met en avant la cacophonie idéologique autour du revenu universel.

Se pose aussi la question de la **politique fiscale**, notamment vis-à-vis des GAFAMI qui sont déjà accusés d'évasion fiscale, avec leur statut d'agent commercial appliqué à leurs filiales. L'IA pourrait accentuer le phénomène de migration de valeur sur les plateformes que ces GAFAMI contrôlent. Cela relance aussi les procédures antitrust en cours, pilotées par l'Union Européenne.

Autre sujet de débat, celui de la **taxation des robots**, proposée notamment par Bill Gates[896]. Pourtant, la taxation des robots est un système bien compliqué, inadapté et n'a pas de sens sans une fiscalité internationale hétérogène. Ce n'est pas plus malin de que de taxer des machines à tisser ou les tableurs Excel !

Si on taxait les robots, il faudrait alors taxer tous les outils matériels et immatériels qui ont amélioré la productivité du travail depuis quatre millénaires : les tracteurs qui ont permis le développement de l'agriculture intensive et de passer, pour prendre la France en exemple, d'une population agraire de 36% des salariés en 1946 à moins de 2% après les années 2000, les logiciels qui ont permis de se passer de secrétaires dans nombre d'entreprises, les machines outils dans les usines, les tableurs qui ont réduit les besoins en comptables, les moteurs de recherche et l'information en ligne qui ont réduit l'attrait des bibliothèques et plein d'autres évolutions du même genre.

Et puis, pourquoi donc taxer les robots physiques alors que l'IA immatérielle pourrait supprimer encore plus d'emplois que les robots logiciels, tout du moins dans les pays développés ? Et si on taxait les robots, cela réduirait l'intérêt économique de rapatrier des usines dans les pays développés, le contraire d'une réindustrialisation. Faudrait-il faire la distinction entre les robots d'usines et les robots humanoïdes ?

La question ne se pose pas pour l'instant. Plutôt qu'inventer une taxe spécifique pour les robots, les Etats pourraient commencer par appliquer sérieusement les taxes génériques qui concernent les entreprises.

Si un Etat met en place une taxation des robots immatériels qui détruisent des emplois, il y en aura toujours d'autres pour accueillir les entreprises concernées et leur servir de paradis fiscal. Il est donc plus important d'homogénéiser la fiscalité que d'en inventer une nouvelle. Si des robots suppriment massivement des emplois, cela améliorera la profitabilité des entreprises et il suffit alors de les taxer correctement sur leurs bénéfices plutôt que sur leur outil de travail qui est une structure de coût et pas de profit.

Une taxe sur les robots appliquée uniquement en France ne ferait que pénaliser l'industrie française par rapport aux autres pays qui font appel à la robotisation, y compris en Asie. Sans compter le fait qu'aujourd'hui, diverses études montrent que les pays les plus robotisés sont ceux qui se développent le mieux !

En taxant simplement les profits, comme on le fait aujourd'hui, on taxe l'ensemble des sources d'économies d'échelle et pas seulement la robotisation. Les entreprises qui sont et seront les plus robotisées auront les meilleurs profits, c'est tout. Il suffit de taxer l'eau qui coule à la fin du circuit économique que dans les multiples robinets qui font tourner l'entreprise.

Il vaut mieux investir dans la formation et les compétences des gens pour les aider à créer des robots, à les installer, les maintenir, les piloter, les superviser, à gérer des projets les intégrant. On ne résiste pas à l'innovation. On s'y adapte et on aide les gens à s'y adapter.

Qui plus est, le risque de pertes d'emplois lié à l'automatisation est plus fort dans les métiers non physiques que dans les métiers physiques. Un expert-comptable est plus menacé par l'IA qu'un kinésithérapeute ou une sage-femme ne le sont par des robots. Dans le cas de la robotisation déma-

[896] Voir Here's how Bill Gates' plan to tax robots could actually happen de Malcolm James, mars 2018, Bill Gates Is Wrong That Robots And Automation Are Killing Jobs de James Bessen, février 2017 et What's Wrong With Bill Gates' Robot Tax de Noah Smith, février 2017.

rialisée, que faudrait-il taxer ? Les logiciels et le cloud ? Ils sont de moins en moins chers. Bref, on tourne en rond.

Enfin, dernier point : l'impact des outils de traduction automatique sur la construction européenne. L'un des écueils de l'Union Européenne est sa grande diversité linguistique qui complique la communication. Est-ce que la traduction automatique va permettre de passer outre ces barrières ? C'est possible. Mais la fragmentation du marché européen n'est pas que linguistique. Elle est aussi forte dans presque tous les secteurs d'activité. Les effets de levier écomique varient d'un pays à l'autre : les médias, les banques, les télécoms, les retailers et les services en général sont le plus souvent locaux. Les USA sont beaucoup plus homogènes de ce point de vue-là et ils sont les principaux pourvoyeurs de produits et services qui sont communs à toute l'Europe, de Google à Amazon.

Souveraineté

Les usages de l'IA posent évidemment des questions clés de ce côté-là.

Sans l'IA, de nombreux services Internet jouent déjà le rôle de régulateurs privés de l'Internet, qu'il s'agisse de Google Search ou de l'algorithme d'alimentation de votre timeline sur Facebook. Avec l'IA, la situation va se corser car les résultats de ces outils vont de moins en moins dépendre d'algorithmes qui peuvent être décortiqués par rétro-ingénierie et de plus en plus de solutions à base de deep learning qui ne sont pas facilement auditables. Les pouvoirs publics en sont à réclamer des solutions techniques permettant d'expliquer les algorithmes. Pourtant, ceux-ci sont compréhensibles par les spécialistes ! C'est le caractère automatisé de la création des « feature maps » intermédiaires des réseaux de neurones convolutionnels qui déroute.

Second enjeu, l'émergence de nouvelles fragilités en termes de sécurité avec la capacité de tromper les techniques de deep learning, notamment dans la reconnaissance d'images ou via les données issues des objets connectées. Cela entraîne un besoin de sécurisation encore plus poussée des infrastructures stratégiques, de défense et de cybersécurité des Etats.

Les Etats devront se doter de leurs propres solutions d'IA spécifiques pour préparer leurs décisions stratégiques et anticiper celle des autres Etats. Cela relève encore de la bordure de la science fiction mais beaucoup moins que les thèses de la singularité.

Les outils du renseignement et de la société de surveillance exploiteront de plus en plus l'IA, notamment pour faire des recoupements d'information pour identifier des profils suspects d'Intenautes.

L'ancien Directeur de la NSA et de la CIA pendant la présidence Bush 43, Michael Hayden, évoquait en 2016 la séparation juridique sur la vie privée et de la sécurité en Europe[897]. Elle est gérée de manière globale aux USA, tandis que dans l'Union Européenne, la vie privée est réglementée au niveau communautaire et la sécurité au niveau des pays. Ce qui créé un handicap pour les Etats.

La souveraineté des Etats sera aussi remise en cause en cas de transformation très radicale de certains métiers. L'impact pourrait être grand dans certains pays où des activités sont délocalisées, comme en Inde. Ces pays seront probablement affectés par la robotisation de ces activités, comme celles qui seront affectées par la RPA (Robotic Process Automation). Dans le même temps, ils pourraient bénéficier des technologies de traduction automatique pour couvrir plus de marchés !

Règlementation

L'IA soulève de nombreuses questions juridiques qui font l'objet de débats depuis plusieurs années et notamment au niveau européen.

[897] Il me semble que c'était dans cette intervention : « Inside the NSA: An Evening with General Michael Hayden », datant de 2014.

En France, le secrétaire d'Etat en charge du numérique, Mounir Mahjoubi, a lancé en juillet 2018 les Etats généraux des nouvelles régulations numériques visant à définir une position française dans les négociations européennes sur les sujets de la régulation numérique et en particulier celles qui sont liées à l'IA. C'est un moyen de formaliser la collecte des retours des entreprises et des startups.

Les retours connus portent surtout sur les critiques concernant les abus de position dominante des GAFA et les blocages opérationnels que peuvent générer l'application du RGPD. Du côté de l'IA, les startups sont surtout en demande de données pour alimenter leurs solutions et en particulier dans le domaine de la santé, et cela ne relève pas forcément des compétences européennes.

La **personnalité juridique des robots** est un autre sujet discuté à l'échelle européenne. L'avocat Alain Bensoussan (*ci-contre*) suggère, avec enthousiasme[898], de créer un véritable droit des robots, situé entre le droit des biens et des personnes[899]. S'y ajouterait la création de référentiels robotiques aux niveaux éthiques, culturels et normatifs. Ce droit comprendrait les règles générales applicables à tous les types de robots, les règles applicables à des robots spécifiques comme les véhicules autonomes, les robots chirugiens autonomes ou les robots de services humanoïdes.

Un robot aurait une identité constituée d'un numéro[900] et même une assurance. Avec l'ambiguité liée au fait que le robot et le logiciel qui l'animent ne sont pas étroitement associés comme dans les êtres humains ou les animaux, le dernier pouvant tourner dans le cloud et servir plusieurs robots à la fois, pouvant aussi être hacké et mis à jour.

Qui plus est, la responsabilité d'un robot en cas d'accident associe son concepteur, le logiciel, les données qui l'alimentent et l'influencent, dont les actions des humains qui l'entourent et les données environnementales. Les responsabilités ne sont plus individuelles, mais des chaines complexes de responsabilités.

200 experts européens ont émis de sérieuses réserves sur la question en avril 2018 via une lettre ouverte[901]. La notion de personnalité juridique du robot a été abandonnée par la commission européenne. Il subistera trois types de responsabilité : celles des concepteurs, celle des sociétés et personnes qui exploitent les robots, et en dernier lieu, celle de ceux qui interagissent avec. Le fonctionnement des robots, et en particulier des véhicules autonomes, devra être auditable pour déterminer les responsabilités en cas d'accidents.

---

[898] Voir son intervention à TEDx Paris en octobre 2015 : De l'urgence d'un droit des robots. Alain Bensoussan a même créé une Association pour le Droit des Robots en 2014.

[899] Alain Bensoussan a aussi lancé divers servcies juridiques en ligne basés sur de l'IA avec sa propre équipe de développeurs. Voir https://www.alain-bensoussan.com/avocat-intelligence-artificielle/. Il propose notamment une solution de justice prédictive.

[900] Mais peut-être aussi accompagné de la version des logiciels qui l'animent, de ses capteurs, de leur état, et des données qui alimentent ses logiciels et peuvent affecter son comportement ! Le robot ne sera pas Skynet mais sa connexion à de nombreux services créera un système fortement maillé difficile à isoler.

[901] Voir Une lettre ouverte pour refuser la « responsabilité juridique » des robots, de Rémy Demichelis, dans Les Echos, avril 2018. Ainsi que « Accorder des droits à une machine, c'est une pente dangereuse » sur la position de Nathalie Nevejans qui est opposée à cette idée de personnalité juridique accordée aux robots, octobre 2018.

Les premiers accidents impliquant des véhicules autonomes ou semi-autonomes ont ainsi permis à un bon niveau de déterminer les fautes entre concepteurs (véhicules n'ayant pas tenu compte de la détection de personnes sur la route comme dans l'accident en Arizona avec un véhicule Uber semi-autonome), conducteurs (n'ayant pas respecté les règles de vigilance) et personnes ayant pu provoquer ces accidents à partir d'autres véhicules (non respect de priorité, …).

La règlementation de l'usage des véhicules autonomes se posera, dont la question de la fameuse gestion des dilemmes lorsqu'un véhicule autonome doit choisir entre deux formes d'accidents et de dommages corporels. Est-ce que la règlementation devra s'appliquer à ce genre de choix moral ? Probablement.

Même si ces choix seront très rares pour ces véhicules. On aura besoin de connaître l'avance le comportement des véhicules pour savoir comment interagir avec. Certains passants s'amusent aujourd'hui à tester la détection de piétons des véhicules autonomes comme les Navya et ont tendance à les empêcher de fonctionner. Un test qu'ils sont moins tentés de faire avec un tram lancé à toute vitesse, sachant qu'il n'aura pas le temps de s'arrêter.

Le second point clé concerne la **protection de la vie privée**, qui risque d'être encore plus mise à mal avec l'IA qui va accumuler et croiser de nombreuses données très personnelles. Avec l'application du RGPD, la réglementation européenne en vigueur depuis mai 2018, les entreprises européennes doivent se conformer à des règles plus strictes sur la protection des données privées. Cela pourrait gêner le déploiement de solutions d'IA grand public et favoriser les GAFA, même si ces derniers devront respecter la même réglementation en Europe. Mais des voix se lèvent aux USA pour réclamer la création d'un RGPD pour les USA[902], y compris celle de Tim Cook, le CEO d'Apple.

Le droit à l'oubli qui est inscrit dans la loi « République numérique » d'octobre 2016 (dite loi « Lemaire ») devra donc s'appliquer aussi aux IA et aux robots de services à qui ont devrait pouvoir demander de ne pas se souvenir d'événements.

On peut se demander comment pourrait fonctionner le droit à l'oubli dans un réseau de neurones complexe dont les paramètres ont été affectés par le comportement d'un utilisateur donné. Faudrait-il réentraîner tout le réseau à partir de zéro pour éviter que celui-ci reconnaisse un utilisateur en fonction de son comportement ?

Quid sinon de l'application du Premier Amendement qui régit la liberté d'expression aux USA, à des robots logiciels ?

Le droit « case law » des USA est très différent du droit romain qui sévit en Europe. Aux USA, une bonne partie du droit provient de la jurisprudence. Il préempte peu l'innovation. En Europe et en France, le droit romain domine et cherche parfois à précéder l'innovation.

Cette différence d'approche a un impact sur la réglementation applicable aux innovations technologiques. Elle favorise plutôt les Américains ! Quand aux Chinois, l'ordre collectif prime sur les libertés individuelles et le droit est donc inversé. La personne se soumet donc tant bien que mal à une société de la surveillance, et notamment de la vidéosurveillance et de la notation comportementale.

Géopolitique de l'IA

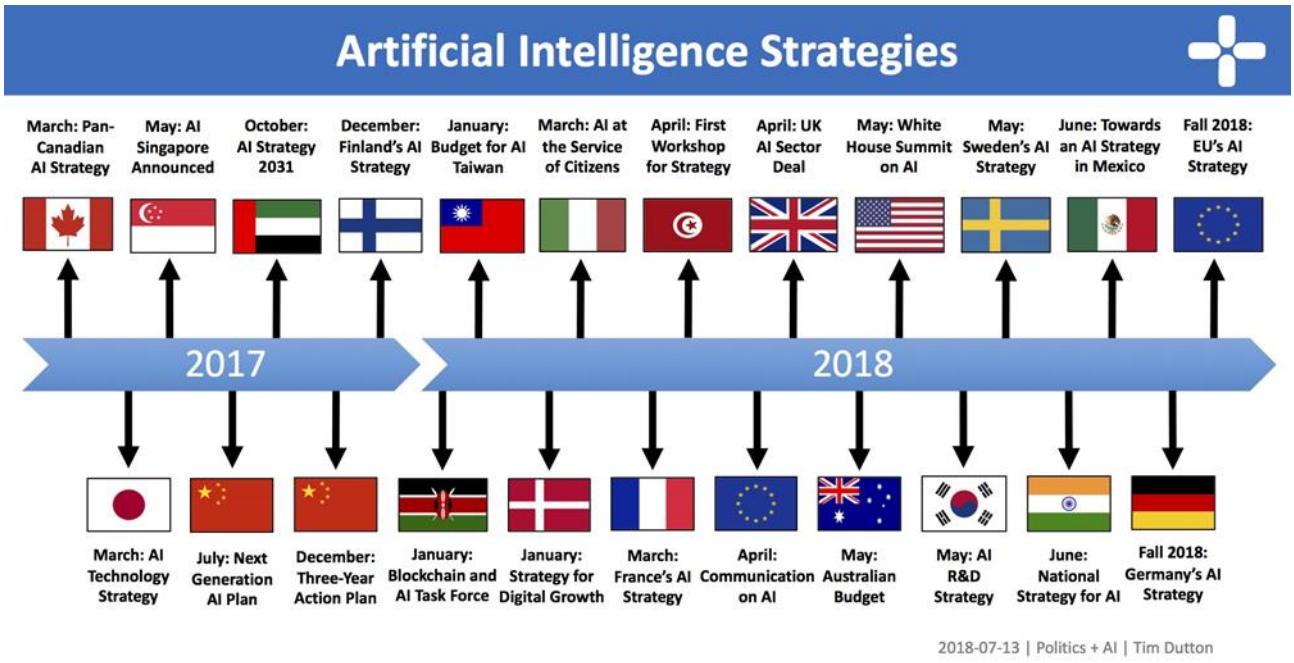
Depuis quelques années, l'IA est devenue un véritable enjeu industriel et politique à l'échelle planétaire. Elle amplifie le jeu de la concurrence entre les trois grands continents technologiques : l'Amérique du Nord, l'Asie et l'Europe.

---

[902] Voir Privacy group calls on US government to adopt universal AI guidelines to protect safety, security and civil liberties, octobre 2018. L'initiative vient de l'Electronic Privacy Information Center (EPIC). Mais rien ne dit que l'Administration Trump en tiendra compte, elle qui passe son temps à détricoter toutes les regulations du marché.

Depuis 2015, tous ces pays, essentiellement ceux de l'OCDE, rivalisent d'annonces et d'investissements pour faire d'eux des « champions » de l'IA[903].

Ils utilisent un panaché des mêmes artifices : des investissements publics dans la formation et la recherche, la stimulation de l'entrepreneuriat, la modernisation de l'Etat aussi bien dans les services publics que dans les activités régaliennes comme celles du renseignement, de la sécurité et de la défense[904].



Les pays se pressent pour devenir leaders de l'IA tout en craignant les effets délétères de l'automatisation sur l'emploi. Dans les pays occidentaux, la question de l'éthique de l'IA est aussi mise en avant. Les ambitions se situent au niveau des zones économiques comme l'Europe, sur les pays, sur les clusters technologiques et sur les capitales. Les pays « tier 2 » comme les pays européens se comparent souvent avec les USA et la Chine. Quantitativement, cela ne peut que leur être défavorable.

Nombre de pays annoncent des montants d'investissements publics dans l'IA. Après en avoir fait le tour, je recommande toujours de les prendre avec des pincettes transparentes. En effet, ces montants sont toujours des compléments à des investissements existants qui ne sont jamais inventoriés. Parfois, ils remplacent des montants existants. Bref, on ne dispose pas toujours d'une vision cohérente de l'ensemble des investissements publics et privés dans l'IA.

Ensuite, les montants sont presque toujours pluriannuels. Il faut les ramener à des montants annuels pour pouvoir les comparer d'un pays à l'autre, puis, les comparer au prorata par rapport au PIB des pays[905]. Le PIB des USA équivaut à 7,5 fois celui de la France !

Comme pour l'entrepreneuriat et les startups, on voit fleurir des classements par pays sur l'IA. Ils sont souvent indexés sur les publications scientifiques et/ou sur les startups d'IA créées[906].

---

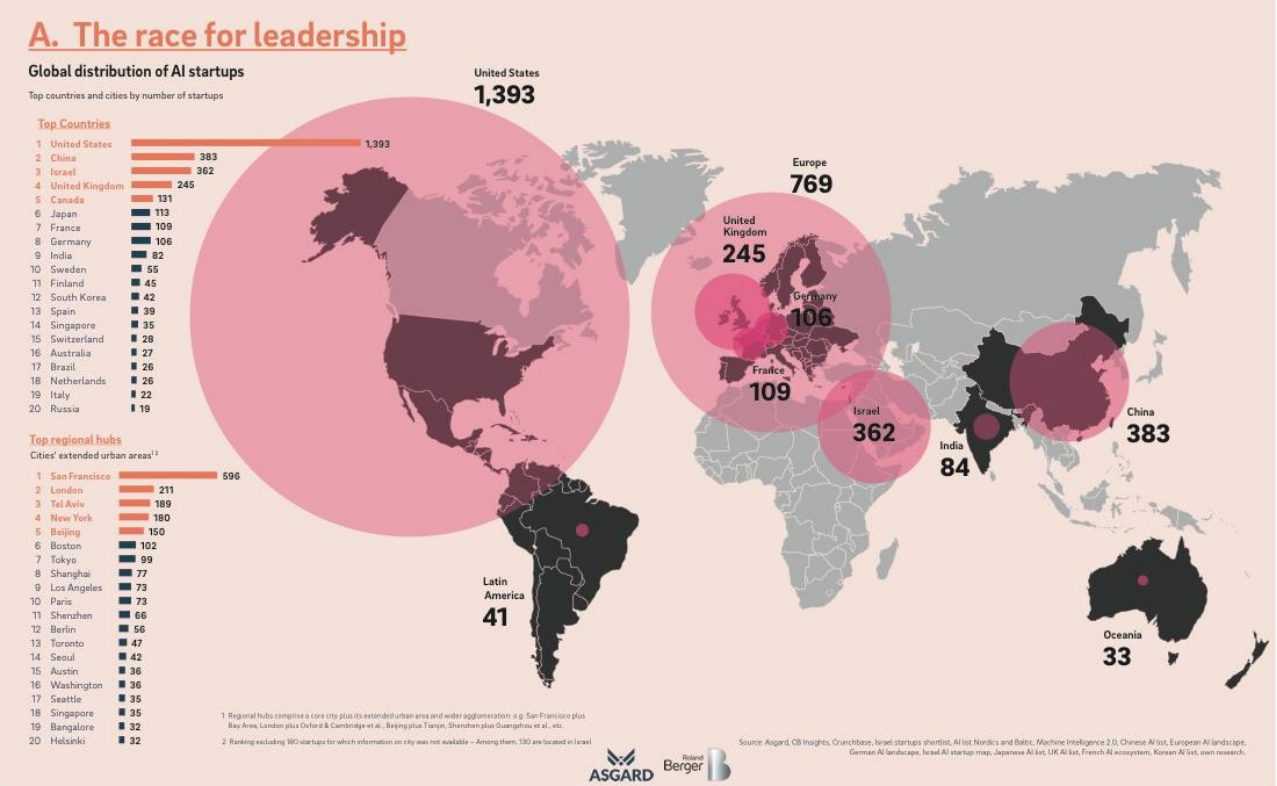
[903] Voir par exemple L'intelligence artificielle, une course mondiale à l'innovation (non daté).

[904] Voir l'inventaire de An overview of national AI strategies, juin 2018 qui couvre 31 pays et leurs initiatives autour de l'IA. Je m'en suis inspiré pour la rédaction de cette partie.

[905] Par exemple, comparer un investissement public annuel d'un pays avec le milliard de dollars qui vient d'être mis de côté pour financer un collège au MIT n'a pas de sens. Ce $1B d'origine privée va alimenter un fonds dont les intérêts annuels serviront à faire tourner ledit collège. Cela donnera donc aux alentours de $30M par an. Si un pays investi un montant voisin chaque année dans un laboratoire ou un établissement d'enseignement public, il sera donc équivalent à ce milliard de dollars !

Le classement *ci-dessous* est originaire d'**Asgard**, une société d'investissement allemande qui propose un inventaire automatisé de startups dans l'IA mais qui a l'air bien incomplet au vu du nombre de startups identifiées rien qu'en France (109).

Il manque dans ces approches au moins la dimension environnementale qui est rarement évoquée. Et puis aussi une pincette de l'impact de cygnes noirs tels que les évolutions politiques de nombreux pays qui tournent vers le populisme[907] et de crises financières cycles inévitables.



USA

Les USA sont actuellement nettement le leader mondial de l'IA en tout cas, dans le mix recherche, industrie et startups. L'écosystème du pays bénéficie notamment de la prodigalité des budgets de la défense et du renseignement qui alimentent nombre de projets de recherche appliquée dans le privé et les universités. Le capital risque est aussi très dynamique pour financer les startups de l'IA.

Enfin, la dominance économique des GAFAMI et leur capacité d'acquisitions de startups leur apporte la puissance industrielle. Celle-ci est aussi soutenue par le rôle que jouent des acteurs tels qu'Intel et surtout Nvidia. Ceci étant, malgré la supériorité économique des GAFAMI, certains Américains s'inquiètent sérieusement de la montée en puissance de la Chine dans l'échiquier mondial. Nous l'évoquerons plus loin.

L'état fédéral US n'a pour l'instant pas officialisé de plan détaillé sur l'intelligence artificielle au-delà de vagues déclarations d'intention. La Maison Blanche avait publié quelques rapports sur l'IA du temps de la présidence de Barack Obama en 2016[908].

---

[906] Voir 5 Countries Leading the Way in AI de Bruno Jacobson, janvier 2018. Il liste cinq pays : Chine, USA, Japon, UK et Allemagne. Ils sont évalués selon leurs publications scientifiques dans l'IA. Sur 20 ans et en 2017, la France est septième selon le Scimago Journal & Country Rank. Voir le classement dans l'IA sur 20 ans http://www.scimagojr.com/countryrank.php?category=1702 et le même sur 2017 : https://www.scimagojr.com/countryrank.php?category=1702&year=2017.

[907] Voir Technological Revolutions Bring About Fascism. Who Will Save Us This Time? de Nicolas Colin dans Forbes, octobre 2018. Il y décrit le lien entre grandes transitions technologiques et émergence du fascisme en faisant un parallèle avec les années 1930.

[908] Voir The Administration's Report on the Future of Artificial Intelligence, octobre 2016, National Artificial Intelligence Research and Development Strategic Plan, octobre 2016 (48 pages) et Artificial Intelligence, Automation, and the Economy, décembre 2016.

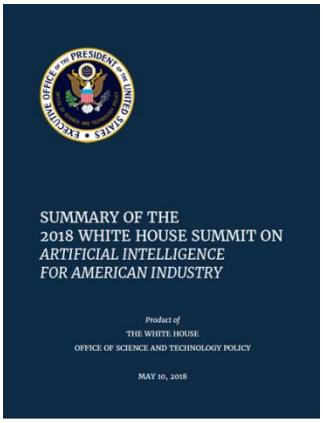
Le plan d'octobre 2016 comprenait le financement de la recherche long terme[909], le développement de méthodes de relations entre humains et IA, le traitement des questions d'éthiques, légales et sociétales de l'IA, assurer la sécurité des solutions d'IA, la création de jeux de données publics pour l'IA, celle d'outils d'évaluation et de benchmarks de solutions d'IA et enfin, l'évaluation les besoins en formation. Sommes toutes des questions assez classiques.

On remarquera qu'aucune mesure ne concernait l'entrepreneuriat et les startups car le système de financement privé US fonctionne bien sans intervention publique directe. Ce plan est passé quasiment inaperçu dans les médias US. Il avait été publié dans la dernière ligne droite de la campagne présidentielle américaine, entre les révélations de l'Access Hollywood Tape de Donald Trump et les déclarations de James Comey relançant pour 10 jours l'enquête sur les mails d'Hillary Clinton.

A partir de janvier 2017, l'administration Trump n'a pas bougé le petit doigt sur le sujet de l'IA pendant un an et demi. Donald Trump a passé plus de temps à relancer (en vain) la production de charbon que sur l'ensemble des technologies et sciences. L'écosystème américain a alors commencé à s'en inquiéter, surtout au vu des avancées de la Chine dans le domaine. Cela devenait un enjeu stratégique et de souveraineté[910].

A la décharge de la Maison Blanche, tant de Barack Obama que de Donald Trump, celle-ci se contente souvent de généralités dans la création de tels plans. Les détails financiers sont laissés à l'initiative du Congrès et la mise en œuvre des plans aux grandes agences fédérales.

Il y a tout de même eu un sommet de l'IA organisé à la Maison Blanche en mai 2018. Le conseiller scientifique adjoint de l'époque, Michael Kratsios, y présentait l'approche dans l'IA de l'administration Trump avec un plan en quatre points comprenant, en particulier, la volonté de supprimer les barrières à l'innovation en assouplissant les règles sur l'expérimentation de véhicules autonomes et de nouvelles méthodes de diagnostic dans la santé[911].



L'un des moteurs fédéraux de l'innovation dans l'IA aux USA est la **DARPA** qui a lancé divers défis tehnologiques comme ceux des véhicules autonomes entre 2005 et 2007 et puis, plus récemment, autour de l'IA explicable (XAI)[912].

Il faut aussi prendre en compte l'action parallèle d'une autre agence fédérale moins connue, l'**IARPA** qui est la DARPA du renseignement, rattachée au Directeur National du Renseignement. Elle cofinance également des projets de recherche fondamentale et appliquée dans les Universités, startups et entreprises, en particulier par le biais de Challenges. Elle est particulièrement intéressée par les applications de l'IA dans le traitement du langage. Par exemple, le Mercury Challenge est un défi pour prédire le futur !

---

[909] Les USA ont investi $1,1B de crédits fédéraux en 2015 dans la recherche civile en IA. Le Pentagone aurait dépensé $7,4B dans l'IA et ses technologies de support comme comme le big data et le cloud pour ce qui est des crédits non classifiés. Ce qui dépend de l'appareil du renseignement (NSA, CIA, DIA, NRO) n'est pas documenté publiquement.

[910] Voir France, China and the EU all have an AI strategy. Shouldn't the US?, de John Delaney, mai 2018, The Trump administration plays catch-up on artificiel intelligence, de Tom Simonite, mai 2018, Trump administration takes a hands-off approach to AI, de Kris Holt, mai 2018 et Hillary Clinton says America is 'totally unprepared' for the impact of AI, de James Vincent, novembre 2017.

[911] Voir Summary of the 2018 White House Summit on Artificial Intelligence for American Industry, mai 2018 (15 pages). Donald Trump n'était pas intervene dans ce sommet.

[912] Voir DARPA announces $2B investment in AI, de Sarah Wells, septembre 2018 qui évoque le lancement du programme AI Next avec un budget de $2B étalé sur 5 ans. Les appels à projet concerneront la sécurité et la résilience des solutions d'IA, la réduction de la consommation d'énergie, les questions de performance et d'explicabilité. En plus de ce budget, il faut ajouter les $1,4B investis sur 5 ans dans le Joint Artificial Intelligence Center qui dépend du DSI du Département de la Défense.

Du côté de la régulation, les USA doivent gérer un délicat équilibre avec d'un côté les lois antitrust qui pourraient amener l'Etat fédéral à limiter la puissance des GAFA et de l'autre, leur intérêt à préserver leur rôle dans l'exercice de la dominance sur l'Internet mondial.

D'un point de vue opérationnel, il faut aussi bien intégrer les mécanismes des GAFAMI lorsqu'ils font leur marketing dans les pays comme la France. Ils investissent localement dans la recherche (Microsoft, Google, Facebook, mais aussi Fujitsu et Naver), dans l'éducation (idem, Cisco, …) et nouent des partenariats avec les grandes entreprises qui se jettent dans la gueule du loup un peu trop facilement alors qu'il ne s'agit souvent que de relations clients/fournisseurs un peu galvaudées et très intéressées[913].

Canada

La recherche canadienne s'appuie sur trois piliers : l'**Alberta Machine Intelligence Institute** (AMII) d'Edmonton, le **Vector Institute** à Toronto et le **MILA** de Montréal où officie Jushua Bengio. C'est à Montréal que s'est installé l'un des laboratoires d'IA de **Facebook** (FAIR) dirigée par Joëlle Pineau ainsi que le laboratoire cortAIx de **Thales**[914].

Les stars canadiennes de la recherche en IA sont **Geoff Hinton**, considéré comme le père du deep learning (maintenant chez Google), et **Yoshua Bengio** (toujours chercheur et ayant résisté aux sirènes des GAFAMI, au MILA de Montréal).

Qu'est-ce qui explique la position du Canada dans l'IA ? En vrac, la qualité de la recherche, des investissements en recherche qui ont été préservés pendant le dernier hiver de l'IA tandis qu'ils diminuaient dans nombre d'autres pays, la bonne collaboration entre les régions même si elles se concurrencent pour obtenir des fonds publics, la création d'instituts spécialisés dans l'IA et une grande liberté dans les directions de recherche choisies. La proximité du marché US est aussi un avantage économique pour les startups. Et on peut remarquer que cette stratégie a aussi fonctionné pour le calcul quantique, autour des laboratoires de recherche et de la startup **D-Wave**.

Le plan **Pan-Canadian Artificial Intelligence Strategy**[915] de cinq ans lancé en 2017 est doté de 82M€. Ils sont investis dans la recherche, ce qui n'est pas grand-chose mais s'additionne probablement à des financements existants. Les axes du plan sont la recherche et la formation supérieure, la création de trois clusters scientifiques et le leadership sur les questions éthiques, économiques, légales de l'IA. Le Canada a aussi lancé l'**Ivado** à Montréal en 2016, l'institut de valorisation des données, qui est un équivalent canadien de nos IRT (Instituts de Recherche Technologiques) associant chercheurs et industriels. Yoshua Bengio est le Directeur Scientifique de l'Ivado.

Cette excellence scientifique du Canada leur a permis d'y voir la création de nombreuses startups, l'une des plus connues était **Element.ai** (2016, $102M). Dans celles que nous avons déjà citées dans ce document, nous avons aussi **Lyrebird** (2017, $120K), **Mindbridge** (2015, $12,7M), **Interset** (2015, $24), **LeddarTech** (2007, $123,3M), **Waste Robotics** (2016), **Analytics 4 Life** (2012, $29M), **Volpara Solution** (2005, $5,5M), **Deep Genomics** (2014, $16,7M), **Finn AI** (2014, $13,7M), **Ross Intelligence** (2014, $13,1M), **Kira Systems** (2015, $50M), **eBrevia** (2012, $4,3M), **Twenty Billion Neurons** (2015, $12,5M), **Rinikloud** (2013, $45,5M), **Granify** (2011, $13,5M), **Landr** (2012, $9M), **Nudge.ai** (2014, $4M) et **DNNresearch** (2012).

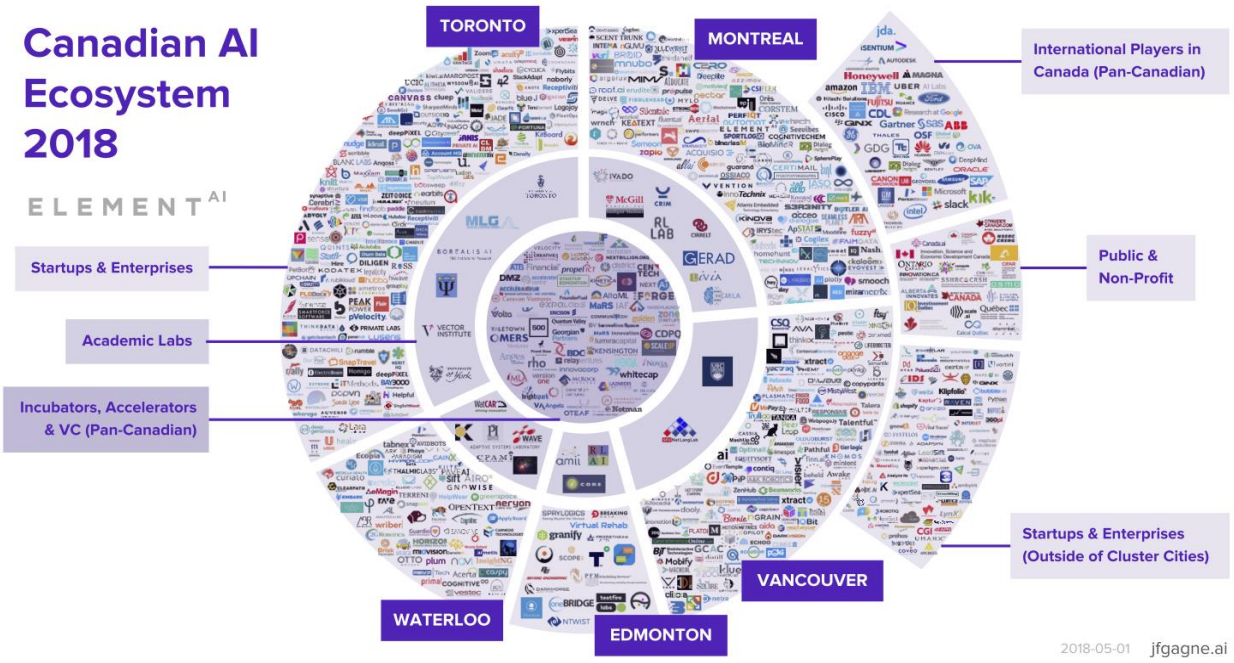
---

[913] Voir Atos et Total s'allient à Google dans l'intelligence artificielle, avril 2018 et Carrefour s'allie avec Google sur l'e-commerce, l'innovation et la bureautique, juin 2018.

[914] Voir Thales choisit le Canada pour son hub mondial en intelligence artificielle, octobre 2017. Basé à Montréal, le laboratoire cortAIx a été lancé en collaboration avec l'Institut Québécois d'Intelligence Artificielle (MILA), l'IVADO (Institute of Data Valorization), l'Institut d'Intelligence Artificielle du Québec et l'Institut Vector de Toronto.

[915] Ce plan est piloté par le CIFAR, le Canadian Institute for Advanced Research, sorte d'équivalent du CNRS. Deux français font partie du conseil scientifique international de l'institut : Yann LeCun (Facebook) et Antoine Petit (CNRS).

Comme l'illustre le graphique *ci-dessous*, l'écosystème canadien de l'IA est dense, surtout comparativement à celui de la France[916]. Ils y présentent les startups et les laboratoires de recherche.



Chine

Est-ce que les Chinois vont dépasser les GAFAMI américains ? C'est une hypothèse à la mode[917]. Les grandes entreprises et startups chinoises du numérique[918] bénéficient d'un terrain favorable : une masse de chercheurs et développeurs formés en Chine ou dans le reste du monde, un marché intérieur de 772 millions d'Internautes (en 2017) et un marché mobile ultra-développé.

Ces grands leaders chinois ont même recruté des talents chez leurs concurrents américains [919]. La recherche chinoise fait des progrès constants même si ils sont moins spectaculaires que ceux des chercheurs US/UK, comme ceux de DeepMind chez Google.

Enfin, les entreprises chinoises bénéficient d'une règlementation qui se pose beaucoup moins de questions sur la protection de la vie privée aussi bien dans la vidéo surveillance qu'en génomique.

---

[916] Source : Canadian AI Ecosystem 2018 (20 slides).

[917] Voir China May Soon Surpass America on the Artificial Intelligence Battlefield, février 2017, China's AI awakening, paru en octobre 2017 et Possible d'être un leader de l'AI d'ici 2030 ? La Chine en pôle position !, avril 2018.

[918] On compte notamment les BATX (Baidu, Alibaba, Tencent, Xiaomi) auxquels il faudrait au minimum ajouter Huaweï qui est le seul des grands chinois massivement présent hors de Chine en plus de Xiaomi. Et pour cause, c'est un fournisseur de technologies, pas un opérateur de services en ligne comme Baidu, Tencent ou Alibaba. Il y a aussi Wechat qui est aussi dans cette catégorie.

[919] Rien que chez Baidu, Andrew Ng qui était auparavant chez Google tandis que Qi Lu provenait de Microsoft et Yahoo. Quand à Hugo Barra, ex Google passé avec fracas chez Xiaomi en 2013, il les a quittés début 2017 et est devenu VP de la réalité virtuelle chez Facebook. Ces prises restent cependant anecdotiques.

Petit à petit, la Chine acquiert son indépendance. On le voit dans les publications de chercheurs sur l'IA, notamment celles que l'on retrouve sur le site **arxiv** de Cornell University. La Chine a largement dépassé les USA en termes de publications scientifiques dans l'IA (*ci-dessous* sur 2017).

Il en va de même pour les dépôts de brevets qui augmentent en flèche depuis 2015 (*ci-dessous*, source CBInsights) et dépassent de loin des USA. La Chine est percue comme ne respectant pas la propriété intellectuelle. C'est le cas de nombreuses PME dans le hardware.

SJR Scimago Journal & Country Rank — Enter Journal Title, ISSN or Publisher Name

Home | Journal Rankings | Country Rankings | Viz Tools | Help | About Us

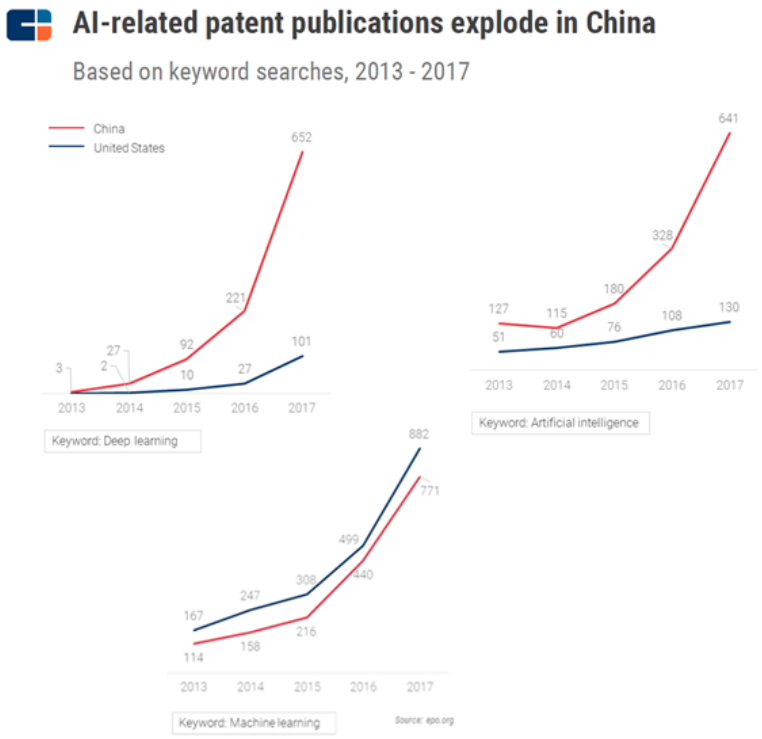
All subject areas | Artificial Intelligence | All regions | 2017

	Country	Documents	Citable documents	Citations	Self-Citations	Citations per Document	H index
1	China	11894	11383	5204	3991	0.44	213
2	United States	8934	8427	3302	1582	0.37	437
3	India	4840	4548	906	433	0.19	108
4	Japan	2985	2847	616	295	0.21	145
5	United Kingdom	2461	2251	1214	439	0.49	228
6	Germany	2212	2018	759	380	0.34	186
7	France	1643	1482	542	216	0.33	162

Mais dans le logiciel et chez les grands acteurs, la Chine affiche une stratégie aggressive de protection et de valorisation de sa propriété intellectuelle.

Mais la dominance de la recherche chinoise en IA est remise en question par les chercheurs occidentaux[920]. La nature des avancées chinoises est à relativiser selon les domaines. Ils sont visiblement meilleurs côté développement d'usages et applications.

Ce qui en soi n'est pas une mauvaise chose puisque cela leur permet d'en tirer un profit économique plus substantiel. Malgré tout, il subsistera encore longtemps un cloisonnement des marchés au niveau des données et des applications entre la Chine et le monde occidental. Sauf accidents de parcours, les grands chinois ne feront pas l'acquisition de GAFAMI.



[920] Voir Is China really that far ahead in AI ? Research says no, de Frank Hersey, août 2018. Avec beaucoup de données sur la bataille des talents entre les USA et la Chine, montrant un avantage des USA, en tout cas en 2017.

Si **Baidu** et d'autres font quelques progrès significatifs dans le deep learning[921], les techniques utilisées qui sont généralement open source ne sont pas des différenciateurs stratégiques suffisants.

Seule la donnée acquise dans leur pays l'est et les données captées par les leaders chinois manquent de diversité pour bien couvrir les besoins à l'échelle mondiale[922]. La majeure partie des grands acteurs de l'Internet chinois n'ont pas de présence dans les pays occidentaux. Wechat, Tencent, Weibo ou Douban sont inconnus chez nous à ce stade de leur développement. Seul Alibaba concurrence dans une moindre mesure Amazon[923].

Par contre, sur certains marchés et notamment dans des technologies enfouies, les Chinois peuvent adopter une stratégie mondiale. C'est le cas de **Baidu** avec son système d'exploitation Appolo OS destiné aux véhicules, notamment autonomes et **Huawei** avec ses processeurs mobiles Kirin 970 et 980 et ses infrastructures télécom. Les startups chinoises sont très focalisées sur la vision artificielle, notamment pour les applications de vidéo surveillance comme avec **SenseTime**, **Yitu**, **CloudWalk** et **Megvii**[924]. Elles ciblent surtout leur marché intérieur.

Les questions de vie privée ne sont pas importantes en Chine au regard de la gestion de « l'harmonie de la société » qui est plus important que les libertés individuelles. Qu'on ne l'aime ou pas, c'est un référent culturel et politique différent des modèles occidentaux.

Les grands acteurs et les startups sont aussi très hardis dans la conception de chipsets d'IA, même si les avancées réelles qu'ils réalisent ne sont pas évidentes par rapport à l'état de l'art US, notamment de Nvidia[925].

La Chine a par contre une forte déficience dans l'influence du monde des développeurs dans l'IA. Très peu de frameworks et outils de développement utilisés en Occident proviennent de Chine aux exceptions près de ceux de Baidu (DuerOS et Apollo OS pour l'automobile). Bref, comparer la Chine aux USA en matière d'IA n'est pas évident. Cela dépend des indicateurs utilisés.

Du côté du gouvernement, la Chine a lancé un **Next Generation Artificial Intelligence Development Plan** en 2016 et annoncé publiquement en juillet 2017[926]. Ce plan aurait été provoqué par le choc de la victoire d'AlphaGo contre Ke Jie, le champion du monde Chinois du jeu, en mai 2017. Un moment « Spoutnik » d'après Kai Fu Lee. Mais la prise de conscience de l'importance avait visiblement démarré bien avant cet événement[927].

---

[921] Voir Baidu spins out its global ad business to sharpen its focus on artificial intelligence de Jon Russel, mai 2018.

[922] Si l'on s'intéresse à la reconnaissance de visages, Google pourrait avoir un avantage sur les leaders chinois. En effet, la qualité d'un tel système est liée à la diversité des visages qui servent à son entraînement. Google est mieux placé pour disposer d'une grande diversité de visages dans ses bases. Il en va de même pour Facebook, ne serait-ce que de par la grande diversité des visages que l'on peut trouver aux USA. Cette diversité est bien moindre en Chine. Pour bien entraîner une IA à base de deep learning, la variété des données fournies est clé ! On pourrait appliquer le même raisonnement à plein d'autres types de données qui sont dépendants de la langue et de la culture : la parole, les modes de consommation, les modes de vie, de transports, etc. Bon, ceci dit, SenseTime est une licorne chinoise qui bat tous les records et sa spécialité est la reconnaissance de visages pour la vidéosurveillance.

[923] Voir A map of the Chinese social media & internet ecosystem, mai 2018 et China Internet Report 2018 (97 pages).

[924] Voir Deep Learning Startups in China : report from the leading edge, juillet 2017 ainsi que Beijing subways may soon get facial recognition and hand scanners, juin 2018.

[925] Voir New Chips, L4 Autonomous Bus & Baidu Brain 3.0 Showcased at Baidu Create 2018, juillet 2018 et Alibaba to launch own AI chip to avoid overdependence on US, septembre 2018. L'activité de chipsets d'Alibaba est liée à l'acquisition du Chinois C-Sky Microsystems en avril 2018.

[926] Voir A Next Generation Artificial Intelligence Development Plan, juillet 2017 et plus de détails dans China's AI Agenda Advances d'Elia Kania, février 2018.

[927] Ceci fait suite à des initiatives favorisant l'entrepreneuriat de masse en 2014 avec la multiplication d'incubateurs et accélérateurs dans les grandes villes, encouragés par l'Etat et financés par les collectivités locales. Il y aurait à la mi 2018 environ 6600 incubateurs de startups en Chine. Les collectivités locales ont aussi créé leurs fonds d'investissements. Le capital risque privé s'est enfin développé par la même occasion.

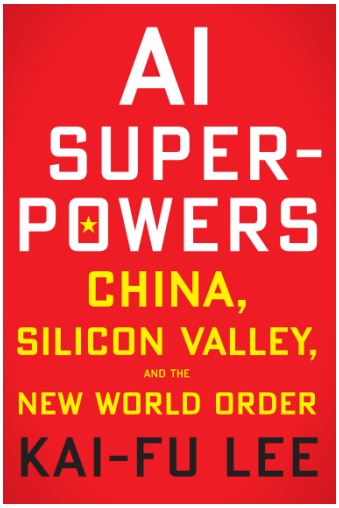
Le plan du gouvernement est complet et couvre la R&D, l'industrialisation et notamment dans le domaine des chipsets, la formation et la capacité à attirer les meilleurs talents mondiaux, l'influence dans la définition de standards et régulations, dans la création de normes éthiques et pour la sécurité. Le plan doit se dérouler en trois phases : un rattrapage là où doit y en avoir d'ici 2020, le leadership dans certains domaines d'ici 2025 et le leadership général d'ici 2030.

Ils ambitionnent de créer une industrie de l'IA de $144B d'ici 2030 et le décuple dans son impact sur les autres industries, dont $1,4T. Pour mémoire, le PIB de la Chine en 2017 était de $12,24T.

Côté thématiques, la Chine se focalise sur les systèmes intelligents de type véhicules autonomes, robots de services, l'automatisation des usines, les systèmes d'authentification et de vidéosurveillance ainsi que dans la santé. Côté hardware, ils visent les capteurs intelligents et les chipsets neuromorphiques. Ils souhaitent en particulier ne pas dépendre des USA sur ce dernier point, surtout au vu de la bataille économique qui fait rage sur les droits de douane, lancée par Donald Trump, sans que l'on sache où cela va s'arrêter.

Le plan comprend le développement d'un parc technologique de recherche en IA de $2,1B AI à Beijing, sur 55 hectares pour attirer 400 entreprises dans les domaines des réseaux très haut débit, du cloud, des systèmes biométriques et du deep learning. Ils prévoient d'y installer la 5G et un supercalculateur. Si la Chine regorge de chercheurs d'excellents niveaux, elle met aussi le paquet dans la formation d'ingénieurs en IA qui vont faire adopter cette dernière par l'ensemble des industries.

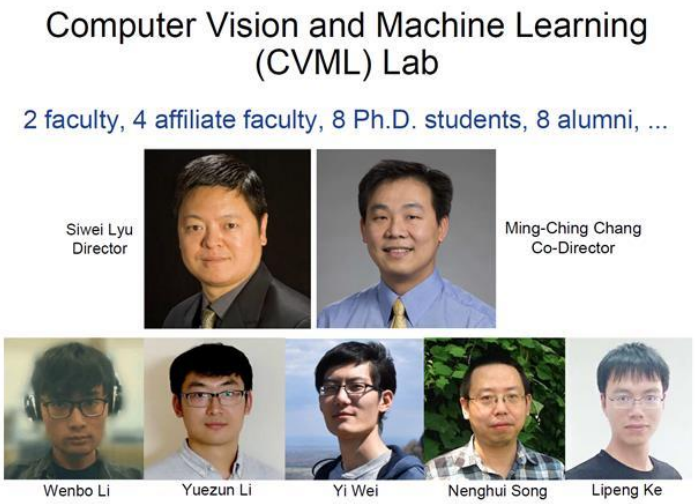
Dans le cadre de ce plan, un nouveau centre de recherché en IA a été créé et inauguré, le **Beijing Frontier International AI Research Institute** en février 2018. Il est dirigé par Kai-Fu Lee qui était successivement chez Apple, Microsoft (MSR China) puis Google. Il avait créé en 2009 le fonds d'investissement **Sinovation Ventures** dédié aux startups en Chine. Il est aussi l'auteur de « AI super powers China, Silicon Valley and the new world order" en 2018[928].



Ce nouveau centre de recherche va entre autres chose fournir des ressources en données et en calcul aux startups en IA chinoises, notamment celles qui sont focalisées sur la vidéosurveillance. Rien qu'à Beijing, il y a plus de 400 startups dont 160 dans l'IA et qui pourront tirer parti de ces nouvelles ressources.

[928] L'ouvrage décrit l'émergence de l'entrepreneuriat Internet en Chine et notamment dans la mobilité. Il illustre comment les startups américaines se sont trompées en abordant le marché chinois, en n'adaptant pas assez leurs offres et applications aux spéficités du marché chinois. Bref, il montre comment les startups chinoises, via une mécanique concurrentielle inouie et darwinienne, ont pu reprendre le contrôle de leur marché intérieur. Kai Fu Lee grossit le trait de la supériorité de l'Internet chinois par le concept O2O, pour online-to-offline, les applications mobiles populaires comme WeChat reliant les utilisateurs mobiles aux services et activités commerçantes du monde physique. La différence chinoise est la concentration de ces activités dans des plateformes mobiles généralistes comme WeChat ou la plateforme d'achats groupés Meituan-Dianping, sorte de Groupon local, qui est notamment devenue très populaire pour la livraison de repas ainsi que pour la location de vélos, via son acquisition de Mobike en 2018. Nombre de ces entreprises n'hésitent pas à gérer la partie physique de leur business alors que leurs homologues américains restent plutôt cantonnés à l'immatériel. Le Uber local, Didi, a une valorisation boursière supérieure à celle d'Uber depuis début 2018 (mais ce sont des valorisations de boîtes non côtées) a ses propres stations services et garages. Le Airbnb local, Tujia, gère ses propres lieux d'hébergement. Kai Fu Lee met en avant le fait que la Chine a plus d'utilisateurs que l'Amérique du Nord et l'Europe réunies. Mais il néglige le décalage de PIB par habitant ($9K/an en Chine, $59K/an aux USA et autour de $42K/an en France/Allemagne/UK). L'approche internationale de ces mêmes startups s'appuie sur l'aide ou l'investissement dans des startups locales, notamment dans les pays émergents, plutôt que sur des déploiements mondiaux à l'américaine. Kai Fu Lee explique aussi pourquoi les USA se font distancer par la Chine dans The US is hastening its own decline in AI, says a top Chinese investor, octobre 2018. Il considère que les US n'investissent pas assez dans des projets ambitieux de recherche pour faire avancer radicalement la portée de l'IA. Il décrit la pénurie de talents alimentées aux USA par d'un côté les GAFAMI et de l'autre une politique migratoire répressive. Par ailleurs, il décrit un futur sombre généré par l'IA, déstablisant l'équilibre social et économique du monde, augmentant les inégalités et le chômage. Au passage, on ne parle pas de GAFA et de BATX en Chine pour ce qui est de l'IA, mais des 7 géants que sont Google, Facebook, Amazon, Microsoft, Baidu, Alibaba, et Tencent. Trois Chinois sur 7 géants ! Il manque juste Huawei pour égaliser. Le livre de Kai Fu Lee évoque les risques que l'IA fait peser sur l'emploi à partir de sa moitié avec une digression sur le sens de sa vie et de l'amour.

Le projet *ci-contre* est-il issu d'un laboratoire de recherche chinois? Non, il provient d'une Université d'Albany aux USA[929]. C'est un scénario que le gouvernement chinois souhaiterait éviter. Autant il apprécie d'attirer en Chine des talents mondiaux, autant il est agacé lors que des Chinois s'expatrient à l'étranger et y restent pour faire de la recherche ou créer des startups. C'est un souci qu'ils ont depuis plus de 10 ans. L'Etat chinois en est au point où il souhaite limiter les échanges d'étudiants et d'ingénieurs avec les USA.

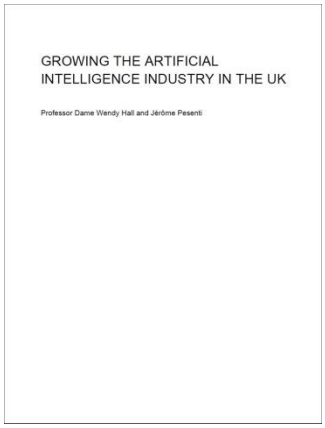


Le marché chinois est sinon toujours aussi protectionniste. Autant les grands acteurs chinois de l'Internet sont absents des marchés occidentaux à de rares exceptions, autant il est difficile pour les leaders américains de s'implanter en Chine. Facebook, YouTube et Twitter n'y sont pas et y sont même interdits. Seul Google reviendrait à la charge avec son moteur de recherche qui respecterait les contraintes du gouvernement chinois.

Royaume Uni

L'approche anglaise est intermédiaire entre celle des USA et de la France pour ce qui est du rôle des pouvoirs publics, avec un interventionnisme mesuré compensé par un entrepreneuriat et un financement des startups plus dynamique.

Côté planification, cela a commencé avec le rapport Growing the Artificial Intelligence Industry in the UK de Wendy Hall et Jérôme Pesenti, en octobre 2017 (78 pages). Il s'agissait d'un rapport indépendant associant une chercheuse et un entrepreneur, en l'occurrence, un Français, ce qui est fort étonnant mais s'explique par la position de Jérôme Pesenti à l'époque[930]. Le rapport préconisait diverses mesures dans l'ouverture des données, dans le développement des compétences, dans l'organisation de la recherche en IA structurée autour d'un UK AI Council, la modernisation de l'Etat avec l'IA. Le Royaume Uni avait déjà créé en 2015 le Alan Turing Institute qui devient des facto le point de coordination de la recherche en IA dans le pays avec un maillage dans les universités et laboratoires de recherche publics.

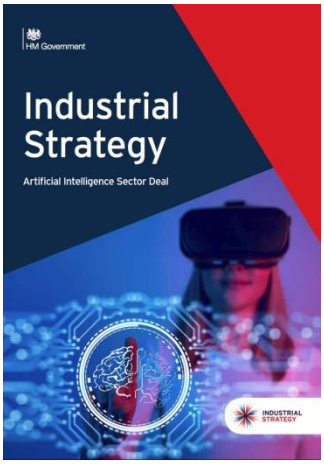


[929] J'ai trouvé ce projet dans ma présentation Video Analytics for AI City Smart Transportation, 2017 (41 slides).

[930] Doctorat en math et en philosophie, Jérôme Pesenti était à l'époque CEO de **Benevolent.ai** (2013, UK, $207M) qui associait l'IA à la médecine pour extraire le savoir des publications scientifiques afin de créer de nouveaux traitements en biotech. Auparavant, il avait créé la startup **Vivisimo** (2000, USA, $5,7M) qui était spécialisée dans l'extraction de données. Après son acquisition par IBM en 2012, il dirige l'équipe technique d'**IBM Watson** dont un tier des effectifs provient de Visimo. Après son passage par Benevolent.ai et depuis début 2018, il est le patron de l'IA chez **Facebook** et devient le patron de Yann LeCun qui gère les laboratoires FAIR (Facebook AI Research). Il partage avec ce dernier le fait d'être un optimiste de l'IA.

S'en suit un autre rapport, produit par la Chambre des Lords, AI in the UK ready, willing and able ?, 2017 (183 pages)[931] et pour lequel Jérôme Pesenti est d'ailleurs auditionné. Dans ce rapport, la France n'est pas du tout citée comme étant un pays investissant significativement dans l'IA. Pour l'Europe, seule l'Allemagne est citée et bien sûr, le Canada, les USA et la Chine. Ce rapport est un peu l'équivalent de celui de l'OPECST français de mars 2017. Il fait le point sur les questions éthiques, sociétales et économiques induites par l'IA. Il s'inquiète notamment des monopoles qui peuvent se constituer autour des données. Il recommande l'audit des données et d'accompagner les PME dans l'adoption de l'IA via un fonds. Il souhaite faire du Royaume-Uni le champion mondial de l'IA éthique !

En avril 2018, le gouvernement et les parties prenantes de la recherche et des entreprises de l'IA publiait Artificial Intelligence Sector Deal (21 pages). Il est cosigné par deux Ministres, celui qui est en charge de l'industrie et l'autre, en charge du numérique et qui a aussi la culture, les médias et le sport dans son portefeuille, ainsi que par la chercheuse Wendy Hall et Jérôme Pesenti, maintenant chez Facebook. Le plan met en regard les investissements publics et ceux des entreprises. Il insiste sur l'investissement au Royaume-Uni d'entreprises étrangères de l'IA venant des USA, du Japon, du Canada et de Hong Kong : Google, Element AI, Amazon, HPE, Beyond Limits, Ironfly, Astroscale et Chrysalix. Les investissements publics spécifiques à l'IA sont de £618M et le privé y met au moins £300M.



Est également créé le **Centre for Data Ethics and Innovation** pour traiter des questions d'éthique de l'IA qui lançait en juin 2018 une consultation publique, close depuis[932].

En mai 2018, Theresa May annonçait un investissement de £210M dans l'IA et la santé, pour permettre aux chercheurs et entrepreneurs d'exploiter les données accumulées par le National Heath Service dans la lutte contre le cancer[933]. Enfin, en juin 2018, le gouvernement UK répondait point par point aux propositions du rapport de la Chambre des Lords d'avril 2018[934]. Le gouvernement insiste beaucoup sur la collaboration avec le privé. Il intervient de manière ponctuelle sans que les investissements soient gigantesques. La forte présence d'entreprises étrangères est un atout pour le gouvernement. Ils l'intègrent d'ailleurs dans la communication de TechNation, l'équivalent britannique de le French Tech depuis 2017.

L'approche du Royaume-Uni est donc un deal entre le public et le privé, ce dernier intégrant même explicitement les entreprises du numériques étrangères, surtout américaines, établies dans le pays.

---

[931] Voir UK can lead the way on ethical AI, says Lords Committee, avril 2018.

[932] Voir Centre for Data Ethics and Innovation Consultation, juin 2018.

[933] Voir Britain to Invest Heavily in AI for Medicine, dans MedGadget, mai 2018.

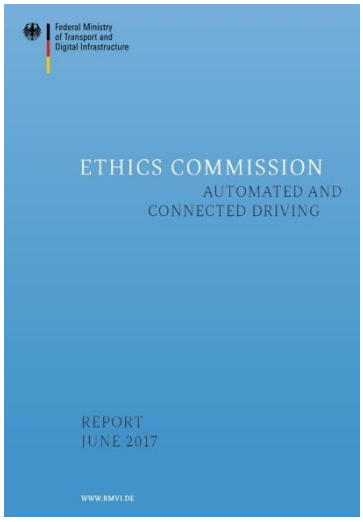
[934] Voir Government responds to report by Lords Select Committee on Artificial Intelligence, juin 2018 et la réponse : Government response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able? (42 pages).

Allemagne

L'Allemagne est assez discrète sur l'IA en tant que telle. Son industrie est mobilisée dans l'Industrie 4.0 qui est à l'industrie allemande ce que la transformation digitale est aux entreprises française, à savoir un grand fourre-tout. Un second domaine est prioritaire : l'industrie automobile.

Le gouvernement présente les premiers éléments de sa stratégie dans l'IA en juillet 2018. Il souhaite accélérer les transferts technologiques entre la recherche et l'industrie. La recherche en IA est concerntrée dans le German Research Centre for AI (DFKI), la Alexander von Humbolt Foundation et les Fraunhofer Institute. Il insiste sur la collaboration européenne et notamment avec la France. On y trouve les éléments habituels comme la formation, l'attraction de talents étrangers, la modernisation des services publics avec l'IA et la promotion d'une IA éthique. Sur ce dernier point, le gouvernement annonçait la création d'une commission dédiée à l'éthique de l'IA avec 19 députés et 19 experts de l'IA. Elle doit créer un set de recommandations d'ici 2020.

Sur un modèle équivalent, le Ministère des Transports et de l'Infrastructure Numérique avait publié un rapport en juin 2017 sur l'éthique des véhicules autonomes[935]. Il traitait des dilemmes dans la prise de décision sur le choix de sauver telles ou telles vies humaines. La réponse ? Il n'y a pas de solution universelle. Ils indiquent que les véhicules autonomes doivent tout faire pour éviter toute forme d'accident. Pour la commission, la morale ne peut pas être intégrée dans les systèmes de conduite. Ils recommandent même en cas d'accident d'éviter de faire quelque discrimination positive ou négative en fonction de l'âge, du genre, de l'état physique des personnes conernées par les dilemmes. Par contre, il serait justifié de minimiser le nombre total de victimes d'un accident donné mais les parties prenantes à l'origine de risques ne doivent pas sacrifier des parties tierces qui n'y sont pour rien.



Ils recommandent une mesure voisine de ce qui se fait dans l'aviation : la création d'un bureau d'analyse des accidents impliquant des véhicules autonomes pour apprendre des accidents et ainsi améliorer les systèmes de conduite autonome.

La situation n'est pas bien brillante côté startups, tout du moins si l'on s'en tient aux levées de fonds[936]. En mai 2019, les plus grandes levées de fonds de startups dans l'IA étaient celles de **Ada Health** ($69M), **Arago** ($55M), **Savedroid** ($53M), **Konux** ($39M), **Wefox** ($34M), **Leverton** ($18), **Zeitgold** ($17M), **Dojo Madness** ($13M), **Door2door** ($11M) et **Fineway** ($11M), sachant que dans le tas, la moitié relèvent de l'IA washing.

Union Européenne

En avril 2018, la Commission Européenne publiait Communication Artificial Intelligence for Europe (20 pages), une première ébauche de plan sur l'IA. Les objectifs affichés étaient de développer la capacité industrielle européenne dans l'IA dans le privé et le public, de se préparer aux boulerversements sociétaux et économiques induits par l'IA et de développer un cadre éthique de l'usage de l'IA. Un investissement de 500M€ était prévu en 2017 et d'un total de 1,5Md€ d'ici 2020[937].

C'était suivi en juin 2018, par la création d'un groupe de réflexion sur l'IA (High-Level Expert Group on Artificial Intelligence) qui doit coordonner les débats de l'European AI Alliance. Ils prépareront des recommendations sur l'éthique de l'IA.

---

[935] Voir Ethics Commission Automated and Connected Driving Report 2017 (36 pages).

[936] Voir Top-10 Artificial Intelligence Startups in Germany, dans CBInsights, mai 2018.

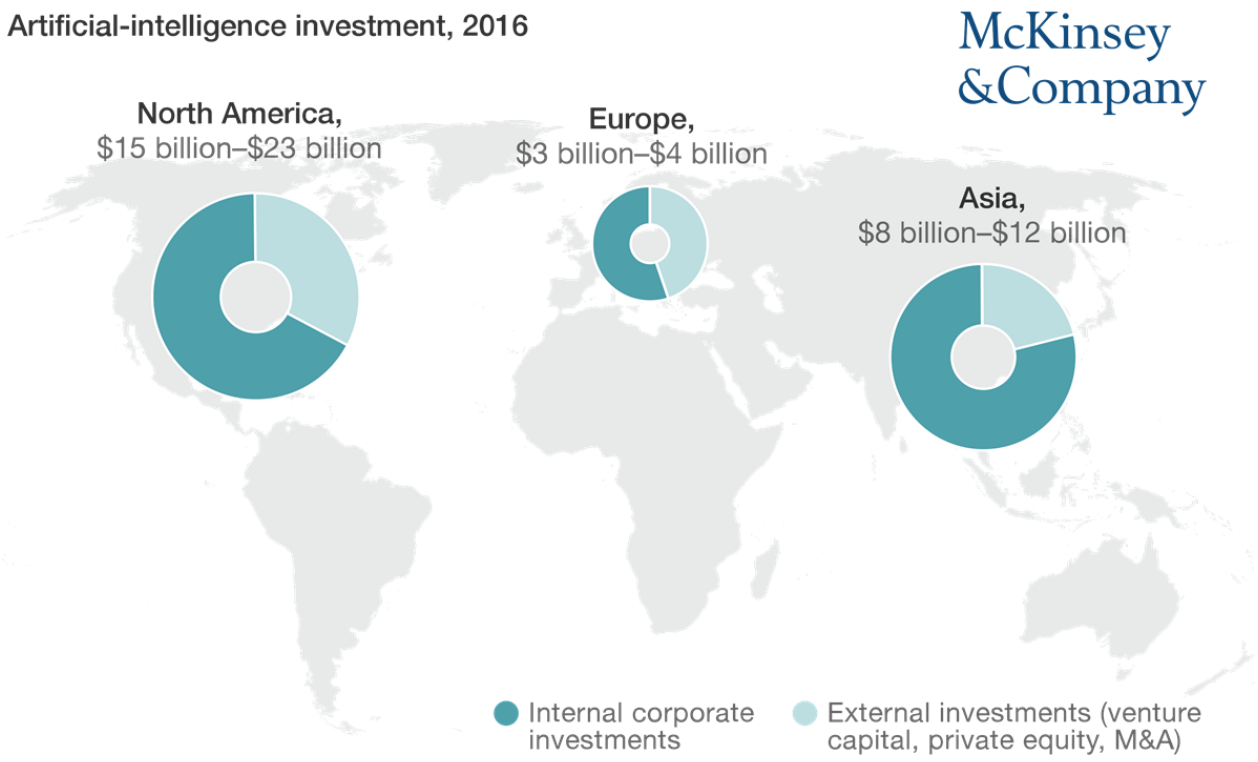
[937] Voir With €1.5 billion for artificial intelligence research, Europe pins hopes on ethics de Tania Rabesandratana, avril 2018.

Le groupe qui est le résultat d'un savant dosage (par pays, genre et métiers) comprend 52 personnes de l'industrie, de la recherche et de la société civile dont 9 Français[938].

La commission planche sur un plan IA qui doit être finalisé fin 2018 avec la volonté habituelle de favoriser les partenariats transnationaux qui conditionnent souvent les financements, notamment ceux d'Horizon 2020 ou des *flagship projects*. Ces projets transnationaux sont poussés par les pays de toutes tailles et par les chasseurs professionnels de subventions. Le problème est que ces projets génèrent beaucoup de friction managériale sans que ce soit forcément justifié par leur taille. Les projets européens ont du sens lorsqu'il s'agit d'atteindre une taille critique comme ce fut le cas avec Airbus ou le CERN.

Et les initiatives ne manquent pas comme ce projet de création d'un institut européen de l'IA sur le modèle du CERN, le **Ellis** Institute (European Lab for Learning and Intelligent Systems)[939]. Ce n'est pour l'instant qu'une proposition de scientifiques, qui vont ensuite s'étriper pour choisir le ou les pays où implanter cet institut s'il est approuvé et financé. La France et l'Allemagne seraient sur les rangs[940].

Il est un secteur où de tels projets pourraient avoir du sens : les semiconducteurs. C'est l'objet du projet **EPI** (European Processor Initiative) visant à créer un processeur européen pour les supercalculateurs mais qui pourrait aussi servir aux véhicules autonomes, une demande des industriels allemands. Il ne sera pas facile à concilier les deux objectifs !



[938] Les 9 représentants français sont Yann Bonnet (ex CNNum et rédaction du rapport de la Mission Villani), Nozha Boujemaa (Inria), Raja Chatila (Sorbonne), Mari-Noëlle Jégo-Laveissière (Orange), Raoult Mallart (Sigfox), Françoise Soulié Fogelman (consultante), Thierry Tingaud (STMicroelectronics), Cécile Wendling (Axa) et Thiébaut Weber (secrétaire général de la confédération européenne des syndicats). Donc, aucune startup, à part Sigfox.

[939] Voir Scientists plan huge European AI hub to compete with US de Ian Sample, avril 2018.

[940] La recherche française signataire de cette lettre comprend Marc Schoenauer (INRIA, et corapporteur du rapport de la Mission Villani) et Stéphane Mallat (Collège de France), Cordelia Schmid (Irnia Grenoble), Francis Bach (Irnia Paris), Jean Ponce (Inria), Julien Mairal (Inria), Jakob Verbeek (Inria), Jean-Philippe Vert (Mines ParisTech et ENS Paris), Emmanuel Dupoux (EHESS, ENS, Paris Sciences Lettre, CNRS, Inria), Theodoros Evgeniou (Insead), Karteek Alahari (Inria), Alexandre d'Aspremont (CNRS, ENS Paris), Pierre-Paul Zalio (ENS Paris-Saclay). On y trouve aussi des Anglais, des Israéliens, des Américains et des Suisses !

Ce projet regroupe 23 partenaires issus de 10 pays vise à assurer une indépendance européenne face aux USA. Le projet est dirigé par Philippe Notton d'**Atos** et ancien de STMicroelectronics. Ce processeur à basse consommation doit servir à créer des supercalculateurs exascale.

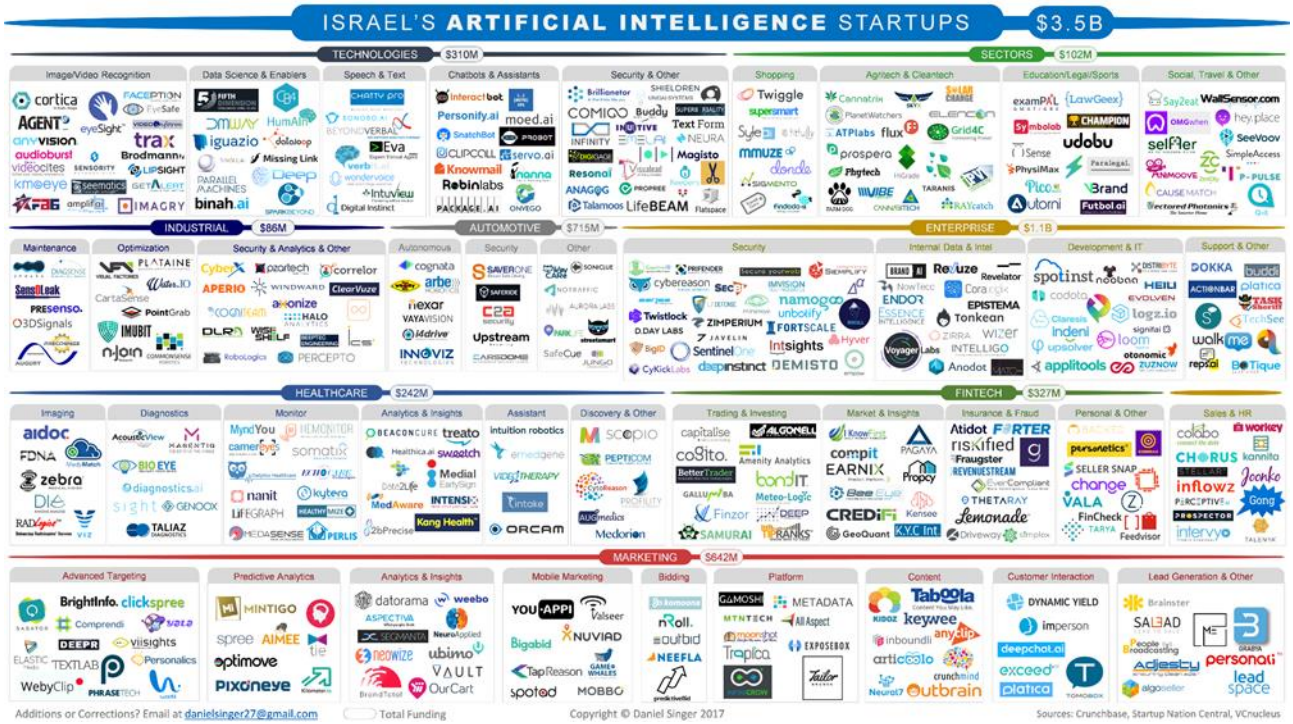
Mais il ne s'agit visiblement pas d'une architecture de tenseur adaptée aux applications de l'IA. Le projet sera « fabless », la fabrication du processeur pouvant être réalisée chez Infineon en Allemagne, STMicroelectronics à Grenoble où en Asie s'il faut descendre en-dessous de 28 nm. Le projet sera financé par l'Europe à hauteur de 120M€[941].

Le projet est un sous-ensemble de l'EuroHPC (European High-Performance Computing) qui vise à créer un supercalculateur européen. Reste à savoir s'il pourra être réellement européen. Quid de la mémoire et du stockage ? Il risque au minimum d'être asiatique.

Du côté entrepreneurial, l'Europe pâtit toujours de la faiblesse de ses acteurs industriels « pure players » du numérique. Les investissements dans l'IA en Europe étaient situés entre 2,4Md€ et 3,2 Md€ en Europe pour une fourchette de 12 à 18Md€ en Amérique du Nord et de 6,5 à 9,7 Md€ en Asie[942]. Le problème des startups européennes est toujours le même : le marché européen est très fragmenté et lent à pénétrer du fait de l'absence d'économies d'échelle et celui des USA conditionne encore nombre des succès du numérique dans le monde, au moins occidental, mais il est couteux à aborder.

Israël

Israël est un OVNI de la sphère mondiale des startups avec une concentration de startups sans égale ramenée au PIB et au nombre d'habitants. Nous n'allons pas refaire l'histoire de l'écosystème entrepreneurial de ce pays qui est très bien décrit dans l'ouvrage **Startup Nation** de Dan Senor et Saul Singer (2009).



[941] Voir European Processor Initiative: consortium to develop Europe's microprocessors for future supercomputers, mars 2018.

[942] Selon McKinsey, dans 10 imperatives for Europe in the age of AI and automation, 2017.

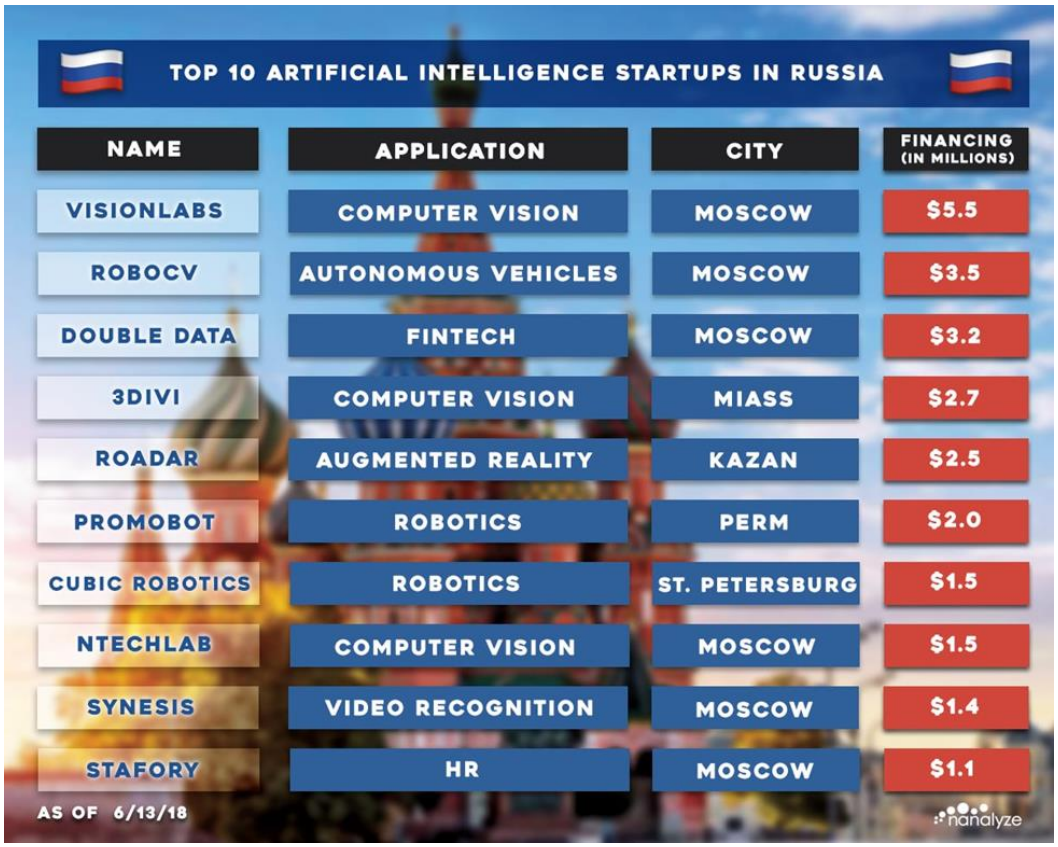
Résultat, dans l'IA, les startups du pays sont abondantes. On en compte presque le double par rapport à la France comme le présente l'inventaire Israel Artificial Intelligence Startups de Daniel Singer publié en septembre 2017. Ces 500 startups avaient levé en tout $3,5B en septembre 2017, dont $837M sur les 9 premiers mois de 2017. Un point clé à noter : 85% d'entre elles sont b2b.

Russie

Vladimir Poutine a contribué à politiser la course à l'IA en septembre 2017 en déclarant que le pays qui deviendrait leader du domaine deviendrait le dirigeant du monde. Il précisait qu'il ne serait pas souhaitable qu'un pays s'arroge un tel contrôle et qu'il faudrait partager le butin. C'est ce que l'on dit lorsque l'on a peu de chances de le récupérer. En effet, c'est un moyen d'admettre que la Russie serait imanquamment sous la houlette des Chinois ou des USA ! Les chances que la Russie joue un rôle de premier plan mondial dans l'IA sont assez maigres au vu de leur production dans la recherche et de leur rythme de création de startups. La vision de Poutine est sinon des plus manichéenne et simpliste.

Les investissements publics de la Russie dans l'IA ne seraient que de $12,5M par an, une bien maigre bouchée de pain[943] ! Et leurs milliardaires prefèrent s'acheter des yachts que de financer la formation des jeunes à l'informatique comme le fait Xavier Niel en France avec l'école 42.

Une bonne partie des investissements russes dans l'IA sont militaires. C'est même le Ministère de la Défense qui donne le la, dans sa liste des 10 priorités publiée en mars 2018[944]. On y trouve du big data, de la formation à l'IA, la création d'un centre national de recherche en IA et l'organisation de « war games » à base d'IA.



[943] C'est un montant étonnament faible. En voici la source : In AI, Russia Is Hustling to Catch Up, avril 2018.

[944] Vu dans Conférence "Intelligence artificielle: problèmes et solutions" – 2018, traduit en français via Google, mars 2018.

Du côté des startups, on constate en effet que les leaders russes n'ont pas fait de mirifiques levées de fonds, témoignant au minimum de leur faible impact international[945]. Leur marché intérieur est insuffisant pour se développer. Le PIB par habitant de la Russie est en effet du niveau de celui du Brésil et au-dessus de celui de la France et du Royaume-Uni, avec des fluctuations dépendant des cours des matières premières.

Le danger provenant de la Russie est surtout géopolitique. Ses services de renseignement n'hésitent en effet pas du tout à hésiter des moyens techniques asymétriques pour faire une nouvelle guerre froide aux démocraties occidentales. Leur ingérence dans l'élection présidentielle US via la combinaison des services actifs du GRU et de la web farm Internet Research Agency financée par un oligarque russe a été dénoncée par plusieurs mises en accusation du procureur spécial Robert Mueller en 2018[946]. Les techniques employées ne faisaient visiblement pas appel à de l'IA contrairement au machine learning mis en œuvre par Cambridge Analytica pour profiler et cibler les électeurs américains. Mais cela pourrait arriver dans les élections suivantes !

Corée du Sud

La Corée du Sud a connu un choc lorsque DeepMind AlphaGo a battu le champion du monde du jeu de Go, le Corée Lee Sedol, en 2016. Deux jours après cette victoire très médiatisée, le gouvernement coréen annonçait un investissement de 770M€ dans la recherche en IA sur quatre ans[947]. Ce plan implique évidemment les grands groupes industriels coréens à commencer par Samsung, LG, Hyundai et aussi Naver. Hyundai est évidemment intéressé par l'IA pour le développement de véhicules autonomes. Samsung a déjà mis sur le marché sont assistant personnel Bixby depuis début 2017. Et Naver investit dans les startups via sa mise de 100M€ dans le fonds d'investissement français Korelya créé par Fleur Pellerin.

L'autre approche de la Corée consiste à installer des laboratoires de recherche à l'étranger. Souvent, les entreprises américaines font cela pour des raisons de lobbying et pour prétendre être bien intérées économiquement avec les grands pays où ils exportent. Ici, il s'agit véritablement de capter des compétences locales. Cela explique l'installation d'un laboratoire de recherche de **Samsung** à New York en robotique ou celui de **Naver** à Grenoble qui y a repris les équipes de recherche de Xerox.

En 2018, le gouvernement coréen lançait un nouveau plan de $1,5B, toujours focalisé sur la R&D. Comme partout ailleurs, le plan vise à accélérer la formation de talents dans l'IA (5000 étudiants, 1400 chercheurs en IA et 3600 data scientists), développement de technologies d'IA dans la santé, la défense et la sécurité, le tout en s'apputant sur un challenge voisin de ceux qui sont lancés par la DARPA.

Le dernier volet est un plan pour développer des semiconducteurs pour l'IA à commencer par un projet d'étude de $28M[948]. Pour créer un chipset d'ici 2029, ce qui semble être une bien lointaine échéance à moins qu'il ne s'agisse de créer une véritable rupture technologique comme avec des memristors, qu'il est très difficile de mettre au point.

Japon

Le pays a été un des premiers à formaliser une stratégie dans l'IA, en avril 2016, sous la forme d'un plan pluriannuel structuré à l'ancienne comme les Japonais en ont le secret (*ci-dessous*).

---

[945] Voir The Top-10 Russian Artificial Intelligence Startups, juin 2018.

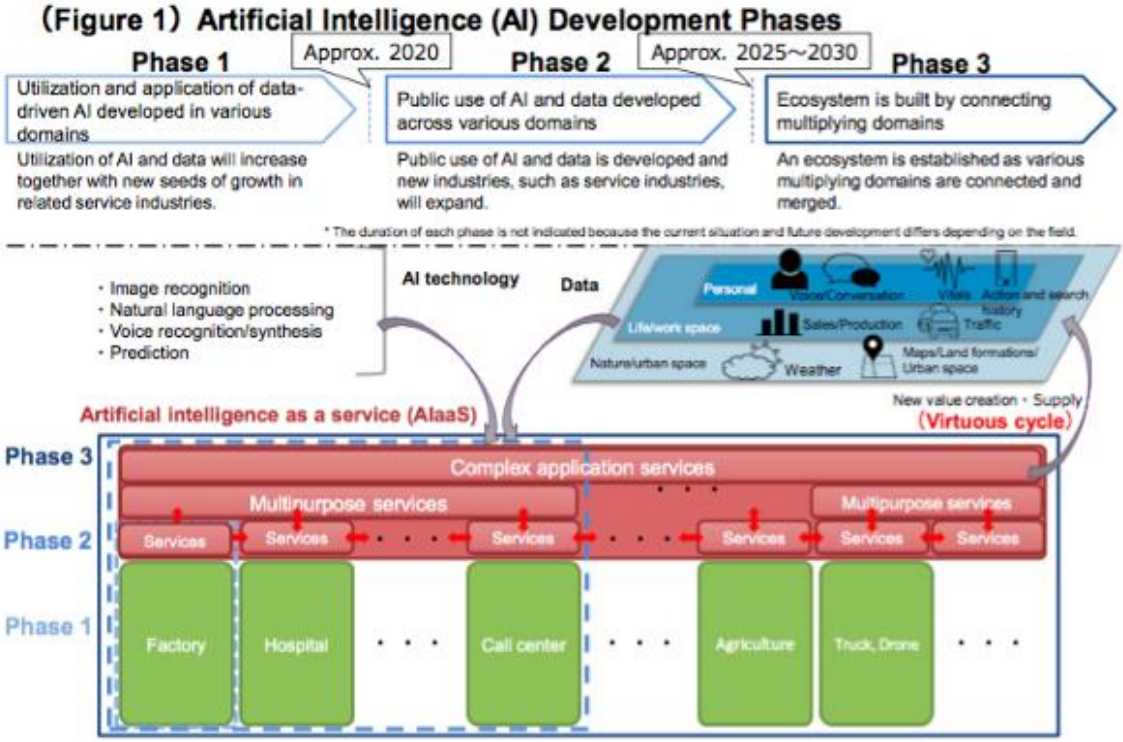
[946] Voir l'inculpation sur la web farm de février 2018 (37 pages) et celle concernant le GRU, juillet 2018 (29 pages).

[947] Voir South Korea trumpets $860-million AI fund after AlphaGo 'shock', mars 2016

[948] Voir South Korean Government to Invest Billions of Dollars into Developing AI Semiconductors, février 2018. D'ailleurs, il n'est pas necessaire d'investir des milliards de dollars pour créer des composants d'IA. En effet, la Corée dispose déjà des capacités de fabrication de semiconducteurs de Samsung. Les besoins en R&D sont donc surtout fabless et peuvent se contenter de quelques dizaines de millions de dollars.

Cela comprend le développement de technologies de base puis d'applications verticales dans les marchés habituels de l'IA (industrie, santé, agriculture, …).

Le plan japonais est formalisé dans Artificial Intelligence Technology Strategy, mars 2017 (25 pages). Ses trois phases sont l'application de l'IA dans divers marchés, l'usage d'IA et de données dans des usages grand public, et la création d'un écosystème multi-discipline. Cela reste très théorique. Il leur reste aussi à dynamiser leur écosystème entrepreneurial, qui est faiblement développé au regard de ceux que l'on trouve en Amérique du Nord, en Europe et même en Chine.



Inde

L'Inde est le troisième pays du monde en termes de publications scientifiques dans l'IA mais le cinquièment en nombre de citations[949]. C'est aussi le second pays du monde en termes de population. Mais il est beaucoup moins développé que la Chine. Le PIB par habitant en Inde est de $1939 pour $8826 en Chine, un rapport de 1 à 4,5 !

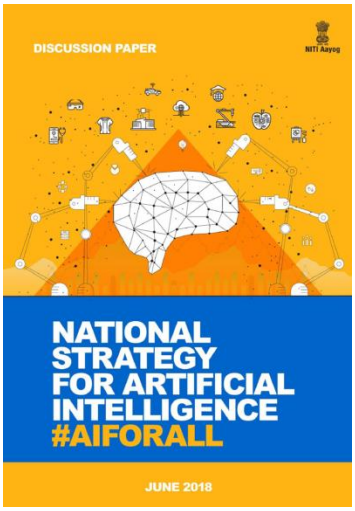
Le pays forme 2,6 millions d'étudiants dans les sciences par an, mais trop dans l'informatique traditionnelle et pas assez en IA.

Le pays a publié sa stratégie dans l'IA en juin 2018[950]. Elle focalise le pays sur ses besoins intérieurs : dans la santé notamment dans les outils de diagostic, dans l'agriculture de précision, dans l'éducation, la smart city et les transports. S'y ajoute le traitement du langage pour la traduction, très utile pour jouer le rôle de passerelle entre les nombreuses langues parlées en Inde, en plus de l'anglais. La situation en Inde est bien différente de la France du côté du marché de l'emploi. En 2016, 51% des emplois étaient dans l'agriculture. Le pays prévoit que cela va descendre à 37% d'ici 2030, soit le niveau d'après guerre en France. Ils prévoient que ce trou d'air sera en grande partie compensé par les emplois dans la construction et dans une moindre mesure dans la santé.

[949] Mais il a bien du mal à transformer cela en innovations et en startups.

[950] Voir National Strategy for Artificial Intelligence, juin 2018 (115 pages). C'est un plan très bien documenté et qui s'appuie beaucoup sur la description de bonnes pratiques internationales, dont la DARPA aux USA et PRAIRIE en France.

Leur activité d'offshore est impactée depuis une dizaine d'années par le déploiement de solutions de RPA (Robotic Process Automation) qui automatise des processus de backoffice de sociétés dans la banque et l'assurance qui faisaient de l'offshore de backoffice en Inde. Il leur faut compenser cela avec des emplois plus qualifiés, d'où les efforts en formation.



Du côté des startups, leur écosystème comprend beaucoup de copycats de projets américains comme **Manthan** ($98M, big data), **Sigtuple** ($25M, vision artificielle), **Haptix** ($12M, chatbot), **Rubique** ($10M, finance), **CreditVidya** ($7M, finance). L'économie numérique du pays dépend surtout des activités de services incarnées par les géants tels que **Wipro** et **Infosys** tout deux situés à Bengalore, et qui pourront développer leurs activités de gestion de projets d'IA b2b qui auront besoin d'autant de prestations de services qu'avant.

Une situation qui fait penser à celle de la France avec ses nombreuses entreprises de services numériques !

Emirats Arabes Unis

Ce petit ensemble de 7 émirats totalisant 9,7 millions d'habitants posé sur un désert en bord de mer n'a pas la prétention de devenir un pôle de recheche et industriel de l'IA. Leur ambition est d'être un utilisateur exemplaire de l'IA dans différents secteurs du gouvernement et de l'économie locale : transports, santé, éducation, espace, énergies renouvelables, eau et environnement.

A l'occasion de la présentation de leur plan stratégique sur l'IA[951], les Emirats Arabes Unis sont les premiers à avoir créé un Ministère de l'Intelligence Artificielle en octobre 2017 dont le premier titulaire, Omar Sultan al-Ulama, n'a que 27 ans. Ils ont d'ailleurs aussi une Ministère de 30 ans en charge des voyages sur Mars qu'ils veulent atteindre avec une sonde inhabitée d'ici 2021.

Iran

De l'autre côté du golfe persique se trouve l'**Iran**. Sa population est l'une des plus éduquées dans l'enseignement supérieur du Moyen-Orient. C'est le 16ième pays du monde en termes de publications scientifiques dans l'IA.

Mais l'isolation économique de l'Iran, tout du moins vis-à-vis des pays occidentaux, s'est agravée avec la sortie des USA des accords du Joint Comprehensive Plan of Action en 2018. Cela impacte par ricochet les partenariats économiques entre l'Europe et ce pays. Résultat, le pays ne peut se tourner que vers l'Asie, qui n'est pas un marché évident à aborder pour eux côté logiciels et Internet, du fait d'acteurs locaux bien implantés comme c'est le cas en Chine.

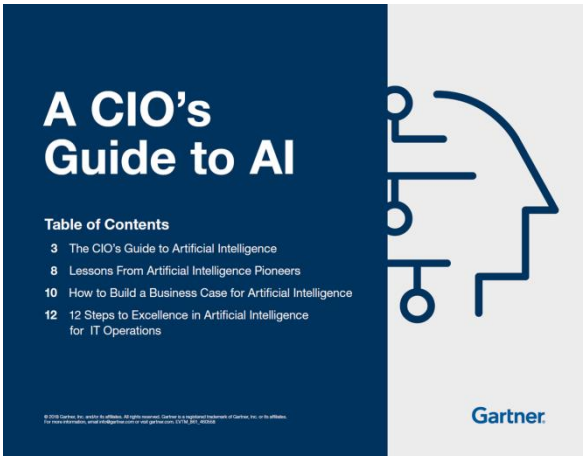
Le monde musulman n'est pas plus facile d'abord car l'Iran représente la minorité chiite face à la dominance du sunnisme. Qui plus est, les exportations du pays dépendent trop du pétrole et du gaz, à 79%.

[951] Voir UAE Strategy for Artificial Intelligence, octobre 2017 créé dans le cadre du plan UAE Centennial 2071.

IA et entreprise

Les entreprises de toutes tailles sont sous le feu de l'incantation de l'inévitabilité de l'IA. Une fois qu'elles ont décidé de faire quelque chose, reste à déterminer quoi, pourquoi, dans quel ordre, comment, avec qui et pour obtenir quels résultats.

Une fois que l'on a découvert le potentiel de l'IA, on découvre que l'on a l'embarras du choix des projets à mener. Il faut gérer les priorités. L'un des critères de sélection des projets est la disponibilité des données nécessaires à l'entraînement des IA.



Contrairement à de nombreux projets numériques, l'adoption de l'IA passe encore plus par des tâtonnements et de l'expérimentation. Les compétences en IA étant rares, les entreprises vont se tourner naturellement vers des spécialistes, de grandes entreprises, des startups et/ou des prestataires de services qu'il faudra choisir sans avoir forcément de repères bien établis[952].

Discours

En à peine trois ans, l'IA est devenue la tendance numéro un du numérique, alimentée par les performances médiatiques des GAFAMI et par l'actualité des startups. L'IA est devenue leur buzzword principal même s'il est maintenant concurrencé par celui de la Blockchain.

L'effet de suivisme est patent chez tous les cabinets de conseils et d'analystes qui ont tous publié leurs livres blancs sur l'IA[953]. Nombre d'entre eux sont lénifiants, rappelant les définitions de l'IA (machine learning, deep learning, vision, langage, …) et présentant quelques études de cas pas toujours bien documentées. Les chatbots ont souvent la part belle dans l'histoire car ils rentrent dans la vulgate de la transformation digitale et de la relation client. Selon le **Gartner**, l'IA était la première des trois grosses tendances de technologies émergentes en 2017 et en 2018[954], avec la réalité mixte et les « plate-formes digitales », c'est-à-dire le reste, avec dans le même sac, la 5G (2020…), les plateformes d'objets connectés, la Blockchain et les ordinateurs quantiques, dont la maturité n'est pas encore au niveau de celle de l'IA.

---

[952] Voir ce guide à l'attention des DSI du Gartner Group, A CIO's Guide to AI, 2018 (13 pages).

[953] Par exemple, le livre blanc proposé dans Intelligence Artificielle : ce qu'il faut savoir pour l'expliquer à son responsable, de MagIT, 2018 (25 pages) est en fait un publi-rédactionnel pour valoriser une étude de cas associant Dataiku et PriceMoo !

[954] Voir Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017, août 2017. C'était toujours le cas en 2018, voir : 5 Trends Emerge in the Gartner Hype Cycle for Emerging Technologies, 2018.

En 2018, l'IA est considérée comme une technologie en pleine phase de démocratisation, même si dans la réalité, cette démocratisation prendra encore quelques années.

Quelques analystes font des prévisions de chiffre d'affaire pour le secteur de l'IA. Certaines se focalisent sur les industries numériques. D'autres, comme sur les objets connectés, portent sur le chiffre d'affaire de l'ensemble des industriels intégrant de l'IA dans leur offre. **PwC** prévoit ainsi que l'IA devrait faire croitre le PIB mondial de $15,7T (trillions, mille milliards) entre 2016 et 2030, soit 14% du PIB mondial actuel, ce qui est probablement exagéré[955]. On nous avait fait le même coup avec les objets connectés ! En 2015, **General Electric** prévoyait que les objets connectés génèreraient $15T de croissance en 20 ans. **Cisco** évaluait cette croissance à $14 ,4T[956] en moins de 10 ans. A chaque fois qu'une évolution radicale du numérique arrive, elle doit ajouter 15 points de PIB. En fait, elle transforme le PIB existant et l'industrie « primaire » du numérique est plus faible que cela. Les loisirs numériques représentent par exemple environ $1T, assez stable dans le temps. On flaire un peu de double booking dans ces prévisions mirobolantes ! Le logiciel influence une bonne partie de l'économie mondiale à lui tout seul.

Voici un exemple de propos marketing bolo-bolo issu d'un acteur des technologies que je ne nommerai pas :

« L'Intelligence Artificielle se développe à un rythme particulièrement soutenu et les entreprises ne peuvent plus ignorer son existence ».

« L'automatisation de tâches simples va permettre aux employés de consacrer plus de temps à des tâches de plus en plus complexes. ».

« Le futur approche à grands pas et les entreprises ne peuvent désormais plus ignorer l'Intelligence Artificielle. C'est aussi simple que cela. L'IA obligera sans cesse les entreprises à se réorganiser, quel que soit le secteur. Il est crucial que les entreprises se penchent sur la question sans attendre. Elles doivent devenir matures sur le plan digital dès maintenant, car elles ne pourront tirer profit de l'IA que si elles ont déjà accompli ces progrès préalables ».

Vous avez aussi **Avanade** dont le rapport **Technology Vision 2017**[957] conseille « *aux entreprises d'intégrer dès maintenant l'intelligence artificielle (IA) pour rester compétitives* ». En précisant que « *les entreprises disposent d'une petite fenêtre de tir pour expérimenter et se familiariser avec les stratégies et les technologies qui préparent à l'arrivée de l'IA dans les pays industrialisés* ». Tout en recommandant aux entreprises « *d'agir avec responsabilité et d'adopter une éthique numérique* ».

S'ensuivent des recommandations qui correspondent probablement aux projets que l'ESN peut mener comme créer des applications avec des interfaces utilisateurs naturelles (vocales, tactiles, VR), des équipes de travail augmentées par l'IA et d'adopter ou de créer des plateformes.

Les discours autour du numérique peuvent être évidemment aussi anxyogènes. C'est le cas de cette étude de **Ricoh Europe** qui agite l'épouvantail de la disparaition face aux PME qui n'innoveraient pas, si possible avec du numérique et un zeste d'IA[958].

---

[955] Voir AI to drive GDP gains of $15.7 trillion with productivity, personalisation improvements, juin 2017.

[956] Voir The Internet of Everything is the New Economy, septembre 2015.

[957] Voir Get ready for the AI first world, 2017.

[958] Voir Sans innovation, un tiers des PME disparaîtra d'ici 2020 d'Ando Randrianarisoa, octobre 2018 dont le titre est évidemment biaisé. Il s'agit d'une enquête terrain d'un fournisseur, Ricoh Europe. Aucun lien sur l'étude n'est fourni dans l'article. Elle est en fait évoquée ici: Failure to innovate threatens to put a third of European SMBs out of business by 2020, octobre 2018, et téléchargeable sur ce lien. En subistance, 34% des patrons de 3300 PME européennes sondées dans 23 pays indiquaient que leur PME pourrait fermer s'ils n'innovaient pas assez vite. Cela ne veut pas dire que cela sera le cas ! Cette étude revient à dire que les PME dont les patrons sont incompétents ont plus de chances de mourir que celles qui ont un dirigeant à la hauteur. En guise de recommandations, l'étude propose aux PME d'améliorer leur relation client, d'utiliser des outils modernes en interne, d'automatiser les processus et de développer la créativité. Evidemment, les solutions de Ricoh sont là pour y pourvoir !

Bref, tous les analystes s'accordent pour dire « *il faut y aller* ». Mais où, avec qui, comment et pour combien, c'est une autre histoire !

Dans « 7 AI myths », Robin Bordoli de la startup **CrowdFlower**[959] (vidéo), synthétise bien les lieux communs sur l'IA que les entreprises doivent comprendre et éviter :

- **L'AI est magique et le deep learning peut résoudre tous les problèmes**. Non. L'IA, ce sont des données d'entraînement, des mathématiques, des probabilités et beaucoup d'itération avec de l'intervention humaine.
- **L'AI est réservée à une élite technologique** et pour les GAFA. Elle est exploitable par toutes les entreprises, notamment via les nombreuses ressources disponibles dans le cloud. Mais il est vrai que l'outillage de l'IA nécessite encore des compétences de bas niveau qui ne sont pas encore suffisamment démocratisées. Les outils d'accès à l'IA vont cependant monter progressivement en niveau d'abstraction pour réduire cette complexité.
- **L'IA est dédiée à la résolution de gros problèmes** valant des milliards d'Euros. Ce document montre qu'il n'en est rien et que les entreprises de tous les secteurs d'activité sont concernées. Dès qu'il y a de la donnée, du langage, de l'image et des règles, l'IA peut faire quelque chose.
- **Les algorithmes sont plus importants que les données**. L'expérience montre que les deux sont aussi importants l'un que l'autre, sans compter le rôle des processeurs. La performance des algorithmes joue un rôle clé dans la qualité des résultats dans le deep learning, et surtout dans la rapidité de la phase d'entraînement des modèles.
- **Les machines sont meilleures que les Hommes**. Non, car les machines ont presque toujours besoin d'interventions et d'informations d'origine humaine. Leur intelligence est alimentée par l'expérience et l'intelligence humaines comme dans les images de radiologie taggées par des radiologues ou des cancérologues. De plus, des IA couplées à des humains sont supérieures aux IA ou des humains seuls. Enfin, les hommes et les machines n'ont pas les mêmes capacités et se complètent.
- **Les machines vont remplacer les Hommes**. En fait, les machines augmentent les capacités humaines et réciproquement. Les métiers qui vont être entièrement remplacés par des IA ou des robots sont rares.
- **L'IA, c'est du machine learning ou du deep learning**. Non. Il existe plein de techniques pour faire de l'IA, notamment autour de l'IA symbolique et du raisonnement automatisé. L'actualité les a mis de côté à cause du raffut autour du deep learning. Mais celui-ci a des limites. Les meilleures solutions d'IA intègrent et assemblent plusieurs techniques différentes.

Méthodes

Les dirigeants doivent aller au-delà du « *j'ai entendu parler de l'IA mais je ne sais pas quoi faire pour l'adopter* »[960].

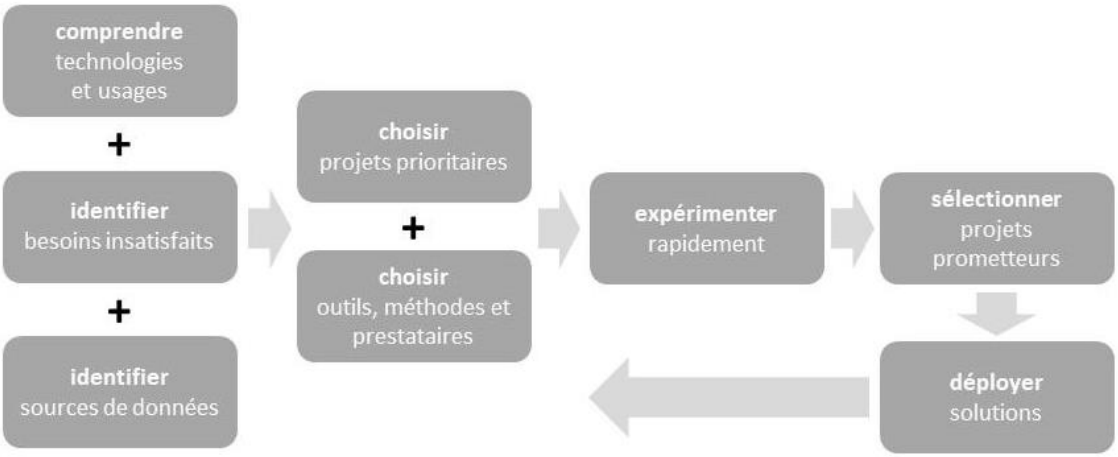
Dans l'IA comme dans de nombreuses nouveaux vagues technologiques, l'innovation va passer par le croisement d'une analyse de besoins mal traités, des attentes des clients, des inefficiences connues de l'organisation, et des potentialités technologiques à portée de main.

Il faut donc avoir comme bagage de départ une certaine compréhension de ce que les différentes briques et outils de l'IA pourraient apporter à l'entreprise. Il faut aussi en connaître les contraintes actuelles, techniques et économiques.

---

[959] **CrowdFlower** (2007, USA, $58M) est une startup qui propose des outils d'exploitation des données pour alimenter des solutions de machine learning. Il automatise les business process amont et aval et met les utilisateurs dans la boucle pour affiner les données et les modèles.

[960] Voir 4 tips to stop reading about AI and start doing AI - How to use AI as a tool in your business de Jana Eggers, janvier 2018.



Il faut ensuite évaluer les données disponibles qui pourraient alimenter des solutions d'IA. Leur volume, leur origine et leur qualité jouent un rôle important dans la qualité d'une solution d'IA bâtie avec.

Le tout peut être touillé dans des réunions d'idéation, sur paper boards et autres Post-it avec les parties prenantes. Il vaut mieux avoir d'abord mis les participants à niveau sur ce que l'IA permet de faire.

La stratégie « données » de l'entreprise sera affectée par l'IA. Les applications de l'IA impacteront la stratégie et les méthodes d'acquisition de données via l'IOT ou d'autres moyens et même sur la stratégie d'open data de l'organisation.

Enfin, on fera le tri des projets potentiels pour choisir ceux qui sont les plus pertinents en fonction de grilles de choix classiques (quick win, coût modéré, avantage concurrentiel, …) puis on passera en phase d'expérimentation. La notion de « preuve de concept » (PoC) est particulièrement valable dans l'IA même si on ne passe pas par une startup pour le mener. La raison est que la majeure partie des solutions d'IA ne génèrent pas un résultat déterministe. Il faut les expérimenter pour en évaluer la qualité. Les solutions d'IA génèrent un taux d'erreur qu'il faut faire descendre aussi bas que la technique et l'équation économique le permettent, et dans la mesure du possible, en-dessous du taux d'erreurs humaines habituelles.

A la suite des expérimentations, il y aura du déchet. Sinon, il n'y aurait pas de processus d'innovation à proprement parler. Seuls les PoC réussis donneront lieu à un déploiement. Et il faudra y consacrer des investissements[961]. On rebouclera alors la boucle pour améliorer les projets déployés et découvrir la potentialité de nouvelles technologies d'IA apparues depuis le début du cycle[962].

Les cycles de développement vont aussi évoluer. Dans l'IA, la mise au point des modèles d'entraînement de machine learning et de deep learning génère des allers et retours plus long que la correction de bugs classiques. L'entraînement d'un modèle peut être très long, même avec les batteries de serveurs les plus puissantes. On ne débogue pas de tels modèles de manière aussi interactive que les langages interprétés du web et même que ceux qui sont compilés.

D'un autre côté, l'expression de ces modèles avec les langages de programmation courants tels que Pythoon et R, couplés à des SDK comme TensorFlow est plus concise. Il n'y a rien d'automatique dans tout cela malgré les lieux communs sur l'IA.

---

[961] Voir IA: rajoutez un zéro! de Sylvain Duranton, BCG, octobre 2018 qui insiste sur l'importance des déploiements après les projets pilotes. C'est là que beaucoup d'énergie doit être investie.

[962] J'ai retrouvé a posteriori des traces de cette méthode dans Artificial Intelligence for the Real World de Thomas Davenport et Rajeev Ronanki dans HBR, janvier 2018.

La mise en œuvre de solutions de machine learning requiert d'expérimenter divers modèles de représentation des données, de segmentation, de prévision. Celui du deep learning passe par la définition de modèles en couches empilés dont la forme et le dimensionnement dépend des données à analyser : images, voix et textes[963].

D'autres recommandations vont dans sens de l'action, comme dans Getting Intelligent About Artificial Intelligence:- 6 Ways Executives Can Start de Davia Temin, décembre 2017. La méthode en six étapes comprend : la compréhension des bénéfices de l'IA, l'analyse des best practices du marché, l'assemblage de compétences internes et externes, l'approche transversale dans l'entreprise, l'accompagnement externe et le démarrage exprimental avant l'industrialisation.

Dans son rapport Gouvernance de l'intelligence artificielle dans les grandes entreprises de septembre 2016, le CIGREF prodigue quelques recommandations sur la méthode à employer pour adopter l'IA dans son organisation et que je vais commenter :

- **Affecter un budget dédié à l'IA** : cela peut avoir du sens pour mettre en place les outils génériques utilisés par les premiers projets. C'est donc une optique de mutualisation a priori. Faut-il des budgets pour les projets eux-mêmes ? Je ne le pense pas. C'est le business qui décide des priorités et l'IA est une technique parmi d'autres techniques avec l'IOT, la BlockChain, la mobilité ou le cloud pour réaliser ces projets.
- **Passer à l'internet 4.0** : IA, algorithmes prédictifs : ce sont des buzzwords. Les algorithmes prédictifs font partie des différentes techniques utilisables mais ne sont pas les seules.
- **Engager un roboticien** dans des équipes IT pour passer en 4.0 : si on est dans des métiers « physiques ». Pour la banque, cela a peu de sens, à moins que cela s'applique aux notions de Robotic Process Automation qui sont liées à l'automatisation de processus métiers de cols blancs et de backoffice.
- **Développer des systèmes de Machine Learning** : il faut s'approprier les outils du machine learning et du deep learning pour en tirer le meilleur parti selon les besoins dans des projets. Si on peut s'affranchir de la complexité de ces outils pour bâtir ses solutions, autant en profiter. De plus en plus d'outils vont dans ce sens pour alléger la charge mathématique et algorithmique des concepteurs de solutions de machine learning.
- **Suivre les tutoriels de TensorFlow** : OK car c'est l'outil générique le plus utilisé pour créer des solutions de machine learning et de deep learning. Et il fonctionne en embarqué ou sur serveurs, on premise ou dans le cloud. Mais ce n'est pas le seul. Il y a aussi des outils plus flexibles comme PyTorch. Et on peut s'approprier le machine learning via des outils d'analyse de données adaptés qui ne nécessitent pas forcément de faire de la programmation.
- **Développer la culture des APIs en interne** : oui, et indépendamment de l'IA, histoire de bien décomposer le système d'information en services interopérant, de favoriser la publication et l'usage d'open data, de transformer son activité en plateforme ouverte aux autres entreprises.
- **Sensibiliser les Métiers et Fonctions aux enjeux de l'IA** : en effet, et en leur faisant des propositions, en croisant les capacités d'usage et les besoins des métiers. L'évangélisation interne des métiers aux capacités de l'IA est une phase importante de l'aculturation. Cela donne des idées. Cette aculturation doit aussi gommer une bonne part de la dynamique anxyogène qui entoure l'IA.
- **Développer une communauté autour de l'IA et échanger** : idem, comme sur tous les sujets technologiues du moment, l'IA étant d'ailleurs souvent un outil associé à ces différents domaines.

[963] Voir cet inventaire des changements qui affectent le développement logiciel dans les entreprises : How AI Will Change Software Development And Applications, par Diego Lo Giudice de Forrester, octobre 2016.

- **Supprimer les « points de douleur » dans l'entreprise** : ce n'est pas spécifique à l'IA. C'est une approche d'innovation passant par l'identification de problèmes à résoudre.
- **Créer des boîtes noires logiques qui gardent en mémoire l'IA** et avoir la possibilité de la détruire (d'effacer l'ensemble des parcs) dans un souci de droit à l'oubli. Effacer les données d'un SI n'est jamais un véritable problème. Les conserver en est un ! La mémoire des IA, notamment à base de deep learning, est située dans les données d'entraînement et dans les paramètres des réseaux de neurones entraînés. Il est important de bien conserver les jeux de données d'entraînement, ne serait-ce que pour pouvoir auditer les systèmes qui deviendraient défaillants.

En octobre 2017, le Cercle de l'IA du CIGREF complétait ce premier rapport avec un nouveau document de 36 pages, Intelligence artificielle dans les grandes entreprises, enjeux de mise en œuvre opérationnelle. Il donne la part belle aux leçons tirées de l'expérience de la mise en œuvre de chatbots clients, avec deux exemples : ceux d'Orange et le cas d'usage dans la RH. Il fait de nombreuses recommandations sur leur mise en œuvre, autour des questions pratique des données qui l'alimentent et de l'éthique.

Le document se fait aussi l'avocat d'une IA ouverte, permettant aux entreprises de faire dialoguer leurs IA entre elles, dans une approche voisine de celle de l'open data. L'architecture proposée est voisine de celle de l'open data accessible via des APIs avec des IA gérant leurs propres ilots de données, qui sont exposées via des services d'accès. Reste à déterminer au cas par cas les données transmises entre IA, dans le respect de la vie privée des utilisateurs et des atouts stratégiques des entreprises. La manière dont l'architecture des SI à base d'IAs semble assez traditionnelle dans l'approche. Le principe même d'un « agent » dans l'IA est d'être relativement indépendant et d'évoluer en fonction de son environnement.

Le Rapport du CIGREF évoque aussi la question clé de la collaboration entre les entreprises et la recherche, en la confondant quelque peu avec les startups dont l'objectif est plutôt de créer des produits qui exploitent de la recherche existante.

La vitesse de transfert des travaux des chercheurs vers les applications d'IA est plus rapide qu'auparavant, ne serait-ce que parce que nombre de travaux de recherche s'appuient sur la publication de code exploitant des jeux de données standards (MNIST, ImageNet, WordNet), faciles à reproduire par d'autres développeurs.

En octobre 2018, le Cigref mettait à jour l'ensemble avec L'intelligence artificielle en entreprise (40 pages) en se focalisant sur la question de la gouvernance des données avec des témoignages d'Orange, Engie, Enedis, La Poste, PSA, Saint Gobain, Air France KLM, Hager, Caisse des Dépôts, de Pôle Emploi et de la Sacem. Le rapport propose aussi diverses formes d'organisation pour adopter l'IA dans l'entreprise (communautés d'experts, challenges d'IA, data innovation labs, etc).

Projets

Supposons qu'un besoin soit bien identifié, que les données soient disponibles et que les technologies de l'IA puissent apporter une solution. Une fois ceci qualifié, on peut rentrer en mode projet.

Va-t-on mener plusieurs "proof of concepts" en parallèle avec différents frameworks d'IA tels que ceux qui sont cités auparavant, et comparer ensuite les résultats ? Ce ne serait pas raisonnable et surtout, ce serait trop coûteux. Il vaut mieux se faire conseiller pour choisir les bons outils et ensuite mener son expérimentation.

Un projet d'IA d'entreprise a un petit côté "recherche applicative", qu'il s'agisse d'un chatbot, d'une application industrielle, d'un système de vision artificielle ou d'un outil d'analyse de données pour faire du prédictif.

Pour prendre ce dernier exemple, on ne va pas juste alimenter une bébête de machine learning ou de deep learning avec un tombereau de données et attendre un beau résultat à la sortie d'un tuyau. Il va falloir d'abord extraire et préparer les données, les nettoyer, les filtrer, savoir ne conserver que ce qui est pertinent.

On va ensuite paramétrer les outils de machine learning ou deep learning en fonction des algorithmes à utiliser. Comme nous l'avons vu dans ce document, il n'existe pas une technique unifiée de machine learning ou de deep learning, mais des dizaines de variantes ! Puis on va observer les résultats. Ils ne seront pas forcément probants du premier coup. Il faudra reboucler sur les données et le paramétrage pour affiner le modèle.

Et il faudra aussi bien visualiser les résultats pour qu'ils soient compréhensibles. La partie « dataviz » d'une application d'IA est aussi importante que les algorithmes retenus.

On appréciera alors la qualité des résultats. Les techniques de machine learning et de deep learning génèrent rarement des résultats exacts à 100%. Il y a toujours un taux d'erreur, que l'on minimise avec l'expérience et que l'on cherche à faire descendre en-dessous d'un niveau acceptable. Comme la variété des échanges typiques acceptables dans un chatbot, le taux d'erreurs d'un système de reconnaissance vocale, ou celui de l'identification de pathologies dans de l'imagerie médicale. A ce jour, les solutions les plus avancées dans ce dernier domaine génèrent un taux d'erreur plus faible que celui des spécialistes ! C'est donc acceptable !

Un benchmark pourra éventuellement avoir lieu pour comparer un projet mené en mode "IA" et un projet mené avec des outils traditionnels de data mining. Si ceux-ci peuvent donner des résultats convenables sur des données chiffrées, ils ne sont maintenant pas du tout à la hauteur pour traiter des données images/vidéo/audio, là où le deep learning est devenu indispensable.

L'IA pose une question habituelle : faut-il développer une solution sur mesure où existe-t-elle sur étagère ? Les éditeurs de logiciels d'IA actuels peinent à créer des solutions véritablement génériques, sauf lorsqu'ils créent des outils de développement. Ils proposent souvent leur solution en mode projet ou service outillé. La raison est que leurs outils n'arrivent pas encore à s'adapter aux sources et structures de données utilisées par les clients, qui varient significativement d'un cas à l'autre. On peut prévoir que cela va s'améliorer à terme, témoignant d'une maturation des projets d'IA en entreprise. Mais on risque de rester assez longtemps comme dans la situation du déploiement d'un ERP qui nécessite beaucoup de développements personnalisé[964].

L'un des points clés pour les entreprises est de passer du projet pilote (PoC) à la production au déploiement à grande échelle. Il y a pas mal de perte en ligne, mais c'est normal lorsque l'on est aux débuts d'un cycle d'innovation.

Benchmarks

Le premier élément d'un benchmark consiste à analyser les études de cas du marché qui sont voisines des projets que l'on souhaite lancer. En matière d'IA, il faut être particulièrement vigilant. Nombre d'études de cas mises en avant par des fournisseurs de technologies exagèrent les résultats voire travestissent entièrement la réalité des projets.

Points de vigilance

Quelques points de vigilance sont à observer : qualifier et quantifier les données qui alimentent les systèmes ainsi que leur origine et leur fiabilité, décortiquer les outils logiques, logiciels et matériels utilisés, et analyser les résultats. Enfin, la structure de coût et la durée du projet sont à intégrer dans l'évaluation.

[964] Voir notamment Build, buy, or both? The AI implementation conundrum, de Pedro Alves, octobre 2018.

Il faut aussi avoir une vision globale d'un projet. Par exemple, un chatbot marketing utilisé dans la relation client doit être évalué sur son impact global sur la satisfaction client et pas seulement sur son impact sur le coût du support commercial ou technique.

Dans le contexte d'un projet d'entreprise, un projet d'IA démarre souvent avec des données et si possible avec de gros volumes de données. Le volume et la qualité des données sont clés pour bien entraîner un moteur de deep learning. C'est l'une des raisons de la force des GAFA : ils ont naturellement accès à d'immenses volumes de données liées aux actions des utilisateurs de Google Search, Facebook, iOS, Android, SIRI, Amazon Alexa, etc. Les sociétés qui déploient de gros volumes d'objets connectés ont aussi accès à des données intéressantes à exploiter.

Un benchmark d'entreprise doit donc partir d'un ou de jeux de données dont on veut extraire quelque chose.

Il faut bien évidemment se poser la question de ce que l'on veut en faire. Au départ, on ne sait pas trop. L'entreprise dispose par exemple d'une base de données du comportement de ses clients et voudrait l'utiliser pour identifier les clients à potentiel d'upsell ou de cross-sell (ventes additionnelles), ou au contraire, ceux qui peuvent générer du churn (abandonner l'offre). Elle peut aussi vouloir déterminer les actions à mener pour optimiser un système complexe : client, production, autre.

L'IA peut aussi servir dans tout un tas de domaines : dans la robotique (qui intègre généralement tout un tas de briques technologiques : vision artificielle, mécanique, systèmes experts, etc), dans la relation client, pour créer des solutions de recommandation, pour analyser des tendances, pour analyser l'image de l'entreprise dans les médias et les réseaux sociaux. Etc. Et la gradation est forte entre générique et spécifique dans ces différentes solutions.

Des projets d'IA peuvent se passer de machine learning et de deep learning et s'appuyer sur des connaissances structurées et des moteurs de règles. C'est par exemple le cas pour créer des systèmes d'assistance à la maintenance industrielle. Dès lors que l'on manipule des données très structurées et une architecture de concepts, les outils de deep learning sont inadaptés. On se retrouve ici dans un domaine ancien, qui a connu ses heures de gloire pendant les années 1980, avec LISP et Prolog. Il n'est pas périmé pour autant, malgré tout le tintouin autour du deep learning, présenté à tors comme une sorte de solution universelle des besoins de l'IA. On va alors faire appel à des **BRMS**, des Business Rules Management Systems.

Conceptuellement, pour les entreprises qui disposent de gros volumes de données, l'IA constitue souvent un ensemble de techniques qui complète une longue lignée de technologies : les bases de données, la business intelligence, le big data, les data analytics et la data intelligence. C'est donc une évolution plus qu'une révolution pour elles.

Modèle d'étude de cas

Vu des clients, il est critique d'accéder à des études de cas de fournisseurs, histoire d'évaluer l'intérêt de lancer tel ou tel projet d'IA dans son entreprise. Voici une proposition de modèle de documentation d'étude de cas de projet intégrant de l'IA[965]. C'est un modèle extensif qui sera probablement rarement complètement rempli. Peu d'entreprises ont envie de documenter leurs projets avec ce niveau de détails. Mais ces études de cas peuvent être réalisées par certains éditeurs pour des projets présentés "behind closed doors".

Société cliente

- Secteur d'activité.

[965] Je reprend ici une proposition que j'avais publiée en décembre 2017 dans Modèle d'étude de cas de l'IA.

- Taille de l'entreprise. Bien préciser la taille de l'entité couverte par la solution. "Total" ou "Orange" n'est pas assez précis. On est souvent trompé par les études de cas qui ne précisent pas leur portée dans une très grande entreprise. Très souvent, les projets n'en concernent qu'une toute petite entité.
- Lieu, ce qui intéressant dans le cas de déploiements internationaux.

Solution

- Description métier du besoin et de la solution. Comment faisait-on avant ? Quelles techniques classiques étaient utilisées ? Quels étaient les surcouts engendrés par l'existant ?
- Description technique de la solution. Quelles techniques d'IA intègre-t-elle : de l'IA symbolique (système expert, moteur de règle, logique floue), du machine learning, des réseaux de neurones simples, du deep learning, des réseaux convolutionnels, des réseaux récurrents ou à mémoire, des techniques de traitement du langage.
- Copies d'écrans de la solution, vue de l'utilisateur. L'interface utilisateur d'une solution logicielle est aussi importante que sa fonction !
- Schémas fonctionnels, un diagramme des flux des données avec leurs sources étant indiqué.
- Périmètre de la solution : projet pilote ou déploiement industriel.

Données

- Nature, volume et origine des données d'entraînement puis de production. Quels capteurs les ont générées (logs Internet, objets connectés, …). Quelles données sont d'origine interne et externe à l'entreprise ? Quelles données exploitées relèvent de l'open data.
- Fréquence de la mise à jour opérationnelle des données. Comment le modèle est-il réentraîné avec l'arrivée de nouvelles données ?
- Taux d'erreur mesuré de la solution si applicable. Ce taux est mesuré après l'entraînement du système d'IA si celui-ci utilise du machine learning ou du deep learning.
- Anonymisation des données exploitées le cas échéant. Est-ce que les données qui alimentent le machine learning ou le deep learning sont bien anonymisées. Normalement, c'est toujours le cas.

Fournisseurs

- Technologies. Au sens : logiciels de base (TensorFlow), d'infrastructure (Spark, Hadoop), progiciels divers et autres.
- Prestataires de services. En indiquant leur apport dans le projet.
- Ressources en cloud si pertinent. Et notamment, si des processeurs spécialisés (GPU ou neuromorphiques) sont utilisés, notamment pour l'entraînement d'un modèle de deep learning.

Dates

- Début du projet.
- Date des premiers tests opérationnels. Ce que l'on appelle un "PoC", pour proof of concept.
- Date de la mise en production. Et portée de la mise en production en nombre d'utilisateurs.

Economie

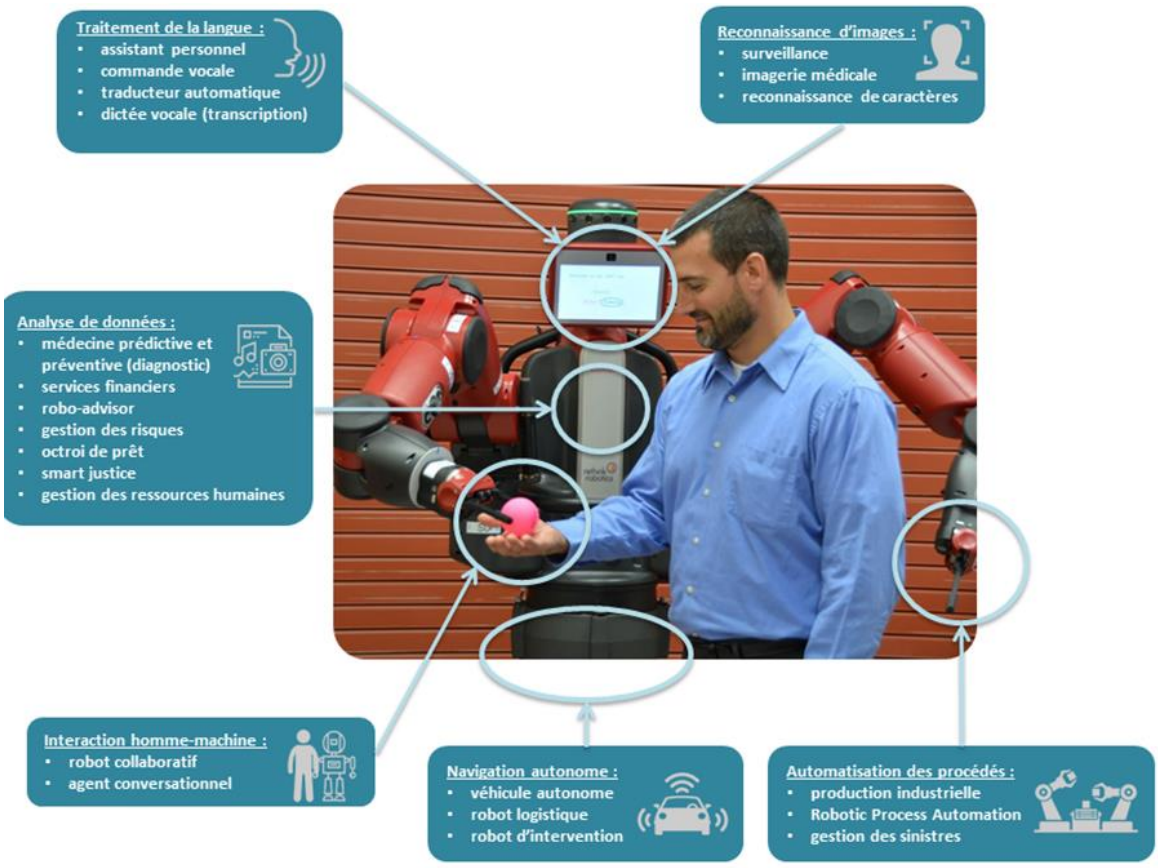
- Coût du projet. Ressources humaines consommées en interne et en externe pour créer la solution. Types de compétences : développeurs, data-scientists, etc.

- Durée d'entraînement des modèles, dans le cas de solutions à base de machine et de deep learning. Et avec combien de serveurs on-premise et dans le cloud et pour quel coût.
- Nombre d'utilisateurs de la solution, aujourd'hui et demain.
- Retour sur investissement. C'est la partie la plus difficile à mesurer sur de nombreux projets. Il faut pouvoir y intégrer l'ensemble des coûts relatifs au projet, y compris la formation des utilisateurs.
- Validation du projet au regard de la RGPD, la Règlementation Générale de la Protection des Données européenne qui entre en vigueur le 25 mai 2018.

Bonne nouvelle, le **LNE** (Laboratoire National d'Essai) s'est lancé en 2018 dans les processus d'évaluation de solutions à base d'IA[966]. Ils proposent cela sous forme de service aux entreprises et couvrent un large spectre de solutions d'IA (*ci-dessous*).

Cela comprend notamment les solutions de traitement de la parole et de reconnaissance d'images. Ils évaluent également le bon fonctionnement de robots, y compris de robots humanoïdes et de robots agricoles, ainsi que celui des capteurs utilisés dans les véhicules autonomes. Cette évaluation est aussi exploitable par les financeurs publics de l'innovation afin d'évaluer les résultats des projets financés de manière rationnelle et normée.

Du côté de l'outillage de benchmarks, on peut citer MLPerf, un outil open source qui permet de gérer ses benchmarks d'applications de machine learning[967]. Il est exploitable pour les solutions de classification d'images, de détection d'objets, de reconnaissance de la parole, de traduction, de recommendation, d'analyse de sentiments et d'apprentissage par renforcement.



[966] Voir Evaluer les Intelligences Artificielle, septembre 2018 et Évaluation des systèmes d'intelligence artificielle.

[967] Voir A new benchmark suite for machine learning, mai 2018 et le site du projet : https://mlperf.org/.

Outils

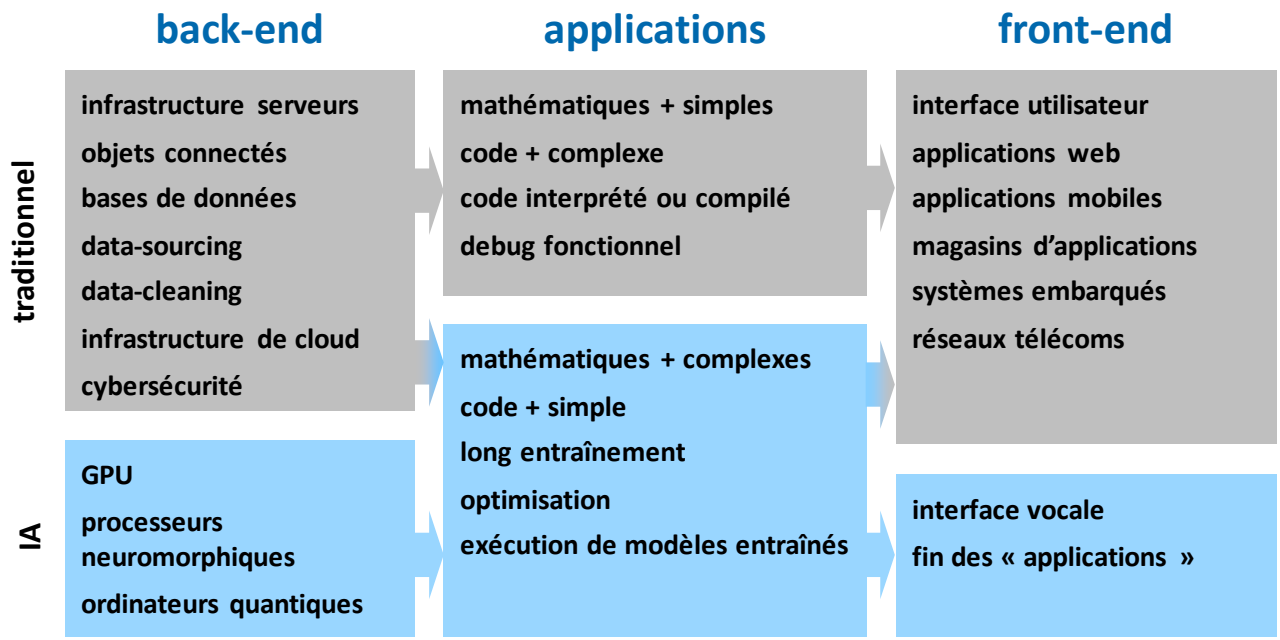
L'entreprise ou ses partenaires devront faire des choix d'outils pour mener leurs projets d'IA. Dans tous les domaines de l'IA, il y a déjà un énorme embarras du choix. La majorité des solutions logicielles de développement sont publiées en open source. Mais il y a des variantes dans leur mise en œuvre. Certains éditeurs proposent ainsi des briques propriétaires autour de leur souche open source.

Les acteurs du marché des outils logiciels de l'IA se rémunèrent avec du service, des solutions métiers payantes, des ressources en cloud, voir du matériel spécialisé.

Il va falloir déterminer où exécuter ses solutions. Si elles sont demandeuses de ressources machine importantes pour l'entraînement ou l'inférence de modèles, avec des serveurs à base de GPU, il sera censé de faire appel à des fournisseurs de telles ressources en cloud. Aujourd'hui, ils doivent être équipés de serveurs DGX-1 ou DGX-2 de Nvidia pour être à la page. Les fournisseurs doivent évidemment être à même de protéger les données de l'entreprise, même les jeux de tests. Ils fournissent des « cloud privés » adaptés à ce genre de besoins.

On fera aussi des choix de topologie d'IA, par exemple, en déterminant où sont réalisés les traitements. Dans certains cas, on les fera réaliser au niveau des capteurs, comme pour certains types de caméras de surveillance qui intègrent leur propres outils de détection d'intrusion et remontent des alertes via le réseau. Certains outils comme la bibliothèque TensorFlow sont conçus pour être exécutés indifféremment sur des objets ou sur des serveurs. Des architectures matérielles comme les GPU de Nvidia le sont tout autant avec des déclinaisons côté serveurs et pour l'embarqué (Jetson)[968].

Néanmoins, les applications de l'IA ne vont pas révolutionner la totalité des systèmes d'information. Elles en exploiteront des briques existantes comme illustré dans le schéma ci-dessous. Ainsi, en amont des outils de machine learning et de deep learning se trouvent des bases de données et des serveurs d'infrastructure traditionnels.



[968] Voir Architecting For AI, de Ann Steffora Mutschler, juillet 2018 qui relate un débat de spécialistes sur le sujet.

La qualité des données alimentant les outils de l'IA. En aval, les applications proposées aux utilisateurs s'appuient toujours sur des interfaces, en général graphiques, adaptées aux micro-ordinateurs, mobiles ou systèmes industriels. L'un des impacts de l'IA est de développer l'usage de la commande vocale.

Il fait même émerger la notion de « VUI » pour Vocal User Interface qui décrit les techniques et bonnes pratiques de gestion de l'interaction vocale avec les utilisateurs. Au passage, les applications vocales intégrées dans les plateformes telles qu'Amazon Alexa sont utilisées de manière transparente par les utilisateurs, faisant disparaître la notion même d'application.

Dans quels cas fera-t-on appel à des chercheurs ? La question se posera lorsque l'état de l'art de l'IA sera insuffisant pour résoudre un problème donné. L'entreprise aura surtout besoin de « recherche appliquée » qui pourra être sous-traitée à un laboratoire de recherche (Inria, un IRT, un 3IA) ou à une équipe de post-docs in situ dans l'entreprise (via contrat CIFRE par exemple). Le point clé à gérer sera l'horizon de temps de ces chercheurs. Il devra être compatible avec celui du projet ! Mais dans la plupart des cas, on n'aura pas besoin de chercheurs, mais plutôt d'une rimbambelle de compétences diverses pour gérer son projet (compétences métiers, data scientist, développeur, designer)[969].

Compétences

Comme pour toute nouvelle technologie, les entreprises font généralement appel à des spécialistes de l'IA qui connaissent la diversité de ses nombreuses techniques et méthodes.

D'après le plan France IA du gouvernement publié en mars 2017, les principaux métiers (et compétence) qui devraient apparaître ou se généraliser autour de l'IA seraient :

- **Architecte en conception d'IA** : une fonction dans la lignée des architectes de systsèmes d'information, qui requiert une vue globale des techniques d'IA et une capacité à les composer, autant dans les architectures logicielles, matérielles que cloud.
- **Intégrateurs d'IA** : il s'agit de développeurs ayant une bonne compréhension et pratique des techniques de machine et de deep learning qui adapteront ces briques technologiques aux usages métiers. Ces développeurs doivent mieux maîtriser les mathématiques des modèles de machine et de deep learning. Ils vont faire plus de maths et moins de code [970] !
- **Spécialistes métier** : que l'on retrouve habituellement dans les fonctions de MOA (Maîtrise d'ouvrage), qui ont une compréhension d'un métier et des données associées, et font le lien entre le besoin métier et les équipes techniques existantes, aident à sélectionner et utiliser les nouveaux outils embarquant une IA.
- **Concepteurs d'interactions avec les IA et robots** : qui maîtrisent l'utilisation de données comportementale et l'ergonomie pour concevoir et spécialiser les interfaces avec les utilisateurs moins qualifiés et les clients. Ils ont des compétences en design !
- **Entraîneurs d'IA** : moins ou pas qualifiés sur les techniques d'IA mais ayant une haute spécialité de leur métier et qui alimenteront en données de référence divers systèmes d'IA. C'est le cas par exemple des systèmes de traitement du langage ou de chatbots qui ont besoin de données textuelles types pour fonctionner.

Dans les petites structures telles que les startups, l'ensemble de ces activités sera concentrée sur un nombre réduit de spécialistes, et même parfois, un seul.

---

[969] Dans Why businesses fail at machine learning, juin 2018, Cassie Kozyrkov de Google explique pourquoi vous n'avez pas forcément besoin de PhD pour votre projet d'IA mais d'un alignement de compétences, surtout métier.

[970] Voir What machine learning means for software development de Ben Lorica et Mike Loukides, juillet 2018.

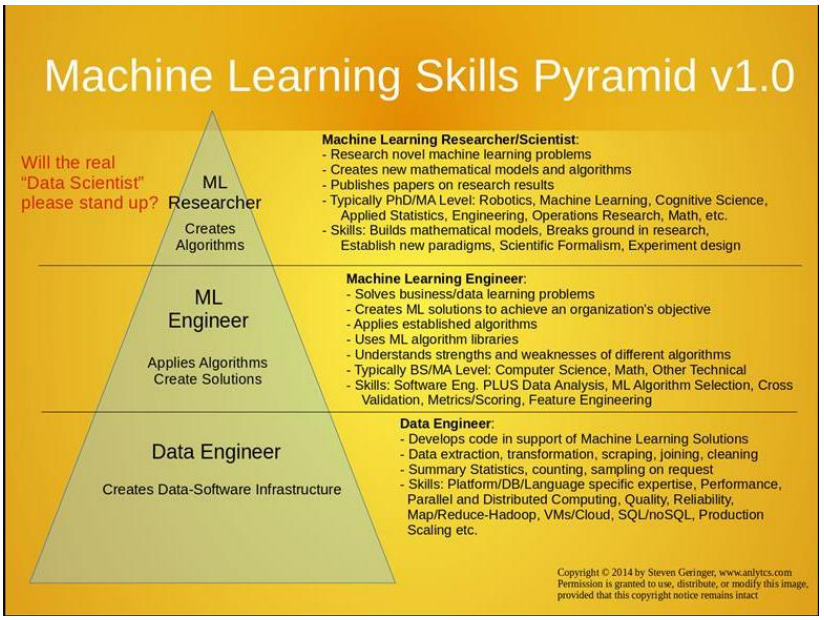
Le marché s'attend également à une forte demande de chefs de projets intervenant de manière transversale sur le développement, l'intégration, et la maintenance des systèmes d'IA, notamment dans les domaines du machine learning, des systèmes experts, du traitement du langage naturel et de la programmation robotique.

Un projet d'IA est comme un projet d'objets connectés : il va devoir réunir des talents et compétences très divers, certaines internes aux entreprises, d'autres externes. La compétence métier prime. Suit la compétence IT plus traditionnelle, pour la collecte et l'exploitation des données.

Le paramétrage des moteurs d'IA passe par des spécialistes d'un nouveau genre qui ont de bonnes bases en IA sachant que la France en forme à peine un millier par an actuellement. Ils sont complétés par des "data scientists" qui jouent parfois tous les rôles.

Après avoir rencontré les pires des difficultés à recruter de bons développeurs, les entreprises de services, les éditeurs de logiciels, les entreprises utilisatrices tout comme les startups vont rencontrer de grandes difficultés à identifier les bons talents à même de paramétrer un moteur de deep learning[971] !

des métiers qui se structurent en permanence

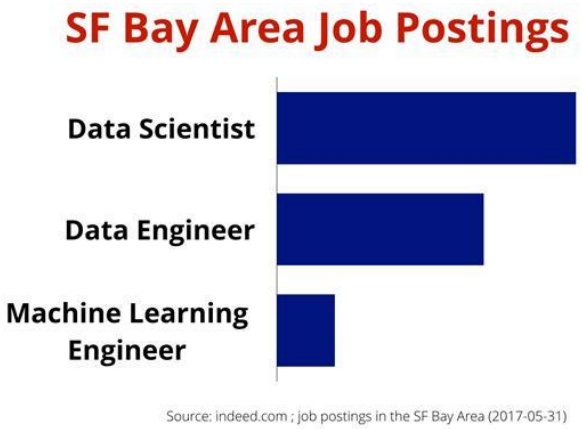


Combien de temps faut-il pour apprendre à paramétrer un réseau de neurones de deep learning ou un système de machine learning ? Il n'existe pas de réponse précise à cette question. Les cursus de formation actuels correspondent à des profils scientifiques BAC+5 avec au moins une à deux années de spécialisation.

On doit pouvoir mettre à niveau de bons profils de développeurs en moins de temps. Ceux-ci ont l'habitude de s'approprier par eux-mêmes de nouvelles techniques et outils. Les plus doués doivent pouvoir s'y mettre de manière expérimentale en quelques mois.

Il faut aussi pouvoir intéresser les salariés des entreprises qui ont été formés initialement à l'IA mais ne l'ont pas mise en œuvre en pratique car ce n'était pas à la mode au moment de leur arrivée sur le marché du travail.

[971] **Kaggle** (2010, USA, $16M), acquis par Google en mars 2017, gère une communauté de data scientists qui lance des défis aux participants. Elle génère des classements qui permettent ensuite d'identifier les meilleurs talents du marché.



Les entreprises de services numériques qui maîtrisent ce genre de projet ne sont pas encore nombreuses. Elles sont en train de s'y mettre. De leur côté, les startups ne sont pas forcément adaptées à la conduite de projets, sauf pour gagner les premiers clients en entreprise. Pour les repérer, on peut commencer par visiter leurs sites web et inventorier leurs représentants, chefs de projets et ingénieurs, qui s'expriment dans les conférences sur l'IA. Les grands acteurs du service vont probablement faire l'acquisition de petits acteurs spécialisés dans l'IA pour étoffer leurs équipes.

Les projets peuvent être vite couteux s'il faut mettre en branle une armée de consultants, data scientists, développeurs et aussi designers. Même si le cœur du réacteur d'un projet d'IA est spécifique à l'IA, autour, il faudra aussi faire tourner des briques plus classiques, tant côté back-end (préparation des données, bases de données, stockage, infrastructure, cloud) que du front-end (créer de belles interfaces pour les utilisateurs).

Où se former à l'IA ? On commence à avoir l'embarras du choix ! Les premières formations sont générales pour faire un tour d'horizon de l'état de l'art de l'IA. C'est ce que je fais chez **CapGemini Institut** depuis fin 2017 (synopsis). Ce même organisme de formation propose un cursus technique sur le machine learning (synopsis). D'autres formations à l'IA sont proposées par la **Cegos** (pour créer un chatbot, synopsis), **Orsys** (big data et intelligence artificielle avec un parcours complet pour les data scientists, synopsis), le **CNAM** (synopsis), **CNRS Formation** (état de l'art de l'IA, synopsis et un catalogue très riche sur le machine learning, le deep learning, le traitement de l'image), **Comundi** (impact de l'IA sur l'entreprise, synopsis), **CentraleSupelec Exed** (enjeux et technologies de l'IA, synopsis), **CapDigital** (avec un parcours pour les PME auquel de je participe pour une courte session sur l'IA, synopsis), l'**Insead** (avec des cursus pour les décideurs, comme AI for business, synopsis), **IA²** (l'Institut d'Automne en Intelligence Artificielle qui propose des formations réalisées par des chercheurs).

Il existe des formations longues comme celles de l'**EISTI** (synopsis) ou de **Simplon.co** (synopsis).

Dans les métiers techniques, on peut aussi choisir de se former par soi-même avec les nombreuses ressources du web. Les cours des grandes universités américains sont souvent en ligne (**Stanford**, **Cornell**, etc). On peut se former en ligne sur les réseaux de neurones et le deep learning sur **Fast.ai**, chercher des cours sur **Coursera** ou sur **DataCamp** (100 cours pour le machine learning et la data science).

Enfin, citons quelques communautés et événements de l'IA avec les **JFPC** (Journées Francophones de la Programmation par Contrainte), Journées d'Intelligence Artificielle Fondamentale, **JFPDA** (Journées Francophones sur la Planification, la Décision et l'Apprentissage pour la conduite de systèmes), **JFRB** (Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes), **JFSMA** (Journées Francophones sur les Systèmes Multi-Agents), les meetups **Machine Learning Applications Group** (Paris, Nantes, Lyon, Pau, Rennes, Aix-Marseille, Bordeaux…). Et puis bien entendu, pour les férus de science, les deux conférences mondiales annuelles de référence, **IJCAI** et **NIPS**.

Organisation

La tentation est grande de créer de nouvelles fonctions de direction autour de l'IA. Après le Chief Digital Officer, le Chief AI Officer ? Pas forcément[972] !

Les équipes existantes peuvent et doivent s'emparer de l'IA :

- Les **DSI** pour intégrer leur dimention technique, l'urbanisation du système d'information, le lien entre les applications de l'IA et le legacy, la gestion des données (data-lake, etc) et les infrastructures.
- Les équipes de **maîtrise d'ouvrage**, où qu'elles soient, pour faire le croisement des potentialités technologiques de l'IA avec les besoins et priorités de l'entreprise.
- Les **CDO** car l'IA est très utile dans les outils associés à leur mission, en particulier dans tout ce qui touche au marketing et à la relation clients.
- Les **business units**, qui doivent être sensibilisées aux bonnes pratiques autour de l'IA dans leur secteur d'activité.
- Les **équipes d'innovation ouverte** qui doivent faire de la veille sur les applications et techniques de l'IA comme d'autres domaines et identifier notamment des startups intéressantes pour les métiers de l'entreprise.
- Les **équipes juridiques** qui doivent être mises dans la boucle lorsque des données personnelles sont en jeu dans les applications de l'IA.

Nommer un Directeur de l'IA serait l'équivalent de nommer un « Directeur du Logiciel » tant l'IA va devenir omniprésente[973]. Ce qui n'empêche pas nombre de grandes entreprises de créer de petites équipes de projets pilotes dans l'IA (chez Thalès, Valéo, Generali, EDF avec son programme Aria, etc) avant que l'IA ne devienne monnaie courante dans les équipes métiers, informatiques et « digitales » (marcom).

Juridique

Le déploiement de solutions d'IA par les entreprises doit être également examiné sous l'angle juridique. Se posent de nombreuses questions liées à la protection de la vie privée que nous avons déjà en partie évoquées ainsi que d'autres qui ont trait à la propriété intellectuelle et au droit des affaires.

Le droit des affaires de l'IA est encore un peu flou. Nombre de juristes travaillent sur l'application du droit existant aux applications de l'IA et à son adaptation éventuelle. Ces débats ont lieu un peu partout dans le monde. Cependant, le droit actuellement en vigueur permet de couvrir un bon nombre de cas d'usages de l'IA.

Voici quelques unes des questions à se poser, sachant que vous ferez appel à vos conseils juridiques et de propriété intellectuels habituels pour en avoir le cœur net !

Responsabilité civile

Les questions en suspens portent sur la responsabilité civile de l'IA. Comment est-elle partagée entre le créateur et l'exploitant ? Cette question va se poser en particulier pour les véhicules et tous les systèmes autonomes.

---

[972] Point de vue partagé dans Qui doit gérer la stratégie en entreprise, de Robin Ferrière, Orange Business Services, septembre 2017.

[973] Voir Qui doit gérer la stratégie d'IA en entreprise, de l'Atelier BNP Paribas, septembre 2017.

Les responsabilités sont multipartites lorsque l'IA est créée par une entreprise et les données qui l'alimentent par une autre, le tout étant exploité par une troisième entité. En cas de problème, on se repose sur les responsabilités humaines qui peuvent être identifiées a posteriori en cas de dfférent juridique[974].

Brevets

Il est difficile de breveter un algorithme en général et celui d'une IA en particulier. Il se trouve par ailleurs que la plupart des algorithmes utilisés dans les applications d'IA sont monnaie courante et pas forcément originaux. Pourtant, des brevets intégrant de l'IA sont régulièrement déposés. Ils utilisent les mêmes contournements que les brevets logiciels en Europe qui ne sont pas possibles sur le papier et sont pourtant nombreux. Les originalités doivent être situées ailleurs que dans les algorithmes à proprement parler.

Quid des inventions créées par les IA ? C'est pour l'instant une question théorique. Le plus souvent, les créations artistiques ou autres créations issues de l'IA sont le résultat de la manipulation d'outils de l'IA par des humains. Ce sont eux qui peuvent obtenir la paternité de ces inventions ou créations au même titre que le graphiste est bien le créateur de ses dessins et pas les outils logiciels qu'il a exploités voir combinés pour les réaliser, aussi sophistiqués soient-ils. Il en va de même pour le créateur industriel qui exploite des outils de conception assistée par ordinateur ou le créateur de musique qui exploite un séquenceur MIDI.

A noter cet avis de l'OMPI ou WIPO, l'organisation mondiale de la propriété intellectuelle qui évoque la création d'outils à base d'IA pour accélérer la validation des demandes de brevets et de marques, notamment pour améliorer les recherches d'antériorité[975].

Secret industriel

Le secret industriel est tout à fait intéressant pour protéger des créations à base d'IA, surtout pour des systèmes d'IA où l'entraînement voir l'exécution des traitements ont lieu de manière protégée dans le cloud. Il est alors difficile d'en faire le *retroengineering* ! Dans ce cas, l'entreprise doit cependant se protéger contre les vols de procédés ou données réalisées par ses propres salariés ou sous-traitants.

Droit d'auteur

Le droit d'auteur s'applique aux logiciels et aux créations de l'IA comme toutes les œuvres déjà créées par l'Homme avec l'assistance de machines. Seule une personne physique est protégée par le droit d'auteur. La machine 100% autonome et créatrice n'existe pas encore[976]. Par contre, on va pouvoir utiliser des IA pour identifier de manière plus rapide d'éventuels plagiats.

Les droits d'auteurs devraient notamment pouvoir s'appliquer à des modèles de deep learning IA entraînés. A tel point que des chercheurs d'IBM ont même développé une méthode de watermarking de modèles de deep learning[977] ! C'est en gros un moyen d'intégrer une signature dans un modèle de deep learning pour vérifier qu'il est bien d'origine lorsqu'il est exploité. C'est une méthode voisine de celle qui protège des contenus protégés par le copyright.

---

[974] Voir Questions juridiques au sujet de l'intelligence artificielle par Marie Soulez du cabinet Lexing Alain Bensoussan Avocats, 2017 (5 pages) et la partie consacrée aux questions juridiques du plan France IA de mars 2017 : Intelligence Artificielle Enjeux Juridiques (19 pages). Insiste notamment sur la notion de responsabilité partagée dans la création de robots entre le matériel et les logiciels, qui ne proviennent pas des mêmes sociétés.

[975] Voir Artificial intelligence and intellectual property: an interview with Francis Gurry, septembre 2018.

[976] Voir La protection par le droit d'auteur des créations générées par intelligence artificielle, un mémoire de Maîtrise en droit de Claudia Gestin-Vilion de l'Université Laval de Québec et de l'Université Paris-Saclay à Sceaux, 2017 (112 pages).

[977] Voir Protecting Intellectual Property of Deep Neural Networks with Watermarking, IBM Research, 2018 (13 pages) ainsi que DeepMarks: A Digital Fingerprinting Framework for Deep Neural Networks, 2018 (12 pages).

Données

A qui appartiennent les données et les modèles entraînés par le machine learning ou le deep learning ? La question est pour l'instant la même qu'avec les données exploitées dans des applications traditionnelles qui n'exploitent pas d'IA.

Contrats

Les entreprises vont faire appel à de nombreux sous-traitants pour créer des solutions d'IA. Quelle est la propriété intellectuelle attachée à ces solutions lorsqu'il y a eu une forte personnalisation de solutions ? Il va falloir faire la part des choses entre contrat de service et contrat de licence d'utilisation de logiciels ou la combinaison des deux.

Epilogue

Nous voici au terme de ce voyage à 360° dans l'IA qui se voulait aussi pratique et pragmatique que possible et destiné avant tout aux entreprises qui se demandent par quel bout prendre le sujet de l'intelligence artificielle.

Vous aurez saisi que les prouesses récentes de l'IA sont liées à des progrès parallèles : dans les méthodes et algorithmes qui s'améliorent continument, dans les données qui les alimentent et dans le matériel. La puissance brute des machines ne fait pas tout, même si elle peut avoir tendance à rendre les développeurs moins astucieux dans leur manière d'aborder les problèmes. Cette impression vient de la difficulté à appréhender la nature même des progrès réalisés dans les algorithmes de l'IA car leur vulgarisation est très difficile. Si vous avez bien saisi comment fonctionnait un réseau de neurones convolutionnel et ses applications, vous avez déjà franchi une belle étape ! Reste à faire de même pour les réseaux de neurones de traitement du langage, ce qui est moins évident.

Vous avez aussi pu constater que l'IA est encore un secteur assez artisanal, aussi bien chez les chercheurs que chez les startups. Elles assemblent des briques techniques de l'IA de manière encore très expérimentale. Les startups du logiciel de l'IA font encore surtout du service outillé. L'évolution de l'industrie du logiciel de l'IA vers une véritable approche produit est semée d'embûches.

Malgré ces nombreux écueils et la bulle médiatique qui l'accompagne, la vague de l'IA est sérieuse et semble aussi importante que les vagues technologiques précédentes qu'ont été le cloud, le big data ou la mobilité. Elle devrait durer. L'IA est en quelque sorte une vague « 2.0 » du logiciel. Elle était en sommeil pendant près de 60 ans, parallèlement avec l'Histoire de l'informatique traditionnelle, et s'est éveillée depuis une dizaine d'années pour ne plus se réendormir. L'IA est une vague technologique aussi importante que l'Internet et l'ensemble de l'économie devra s'y mettre.

Nous avons pu découvrir de nombreuses startups françaises et étrangères, aussi bien au niveau des techniques horizontales que des applications métiers. Nous ne manquons pas de chercheurs et d'entrepreneurs de l'IA en France. Les questions qui se posent sont les mêmes que d'habitude : comment faire en sorte que ces startups grandissent vite, soient bien financées et se développent à l'international. Les startups et grands acteurs du numérique US et chinois ne nous ont pas attendus !

Même avec une IA un peu faiblarde et lourdingue, la marche vers l'automatisation partielle de nombreux métiers est déjà en route et va dans le sens d'une histoire qui a démarré avant l'invention de la roue. L'IA est une nouvelle boîte à outils à intégrer dans son atelier logiciel. Il faut s'y préparer dès maintenant, ne pas y résister futilement, s'y adapter en se modernisant, en faisant évoluer notre système d'enseignement et en produisant des outils compétitifs.

Les civilisations qui ont évité les progrès techniques et les outils de communication dans l'Histoire ont systématiquement périclité ou, au mieux, décliné. Le futur n'est pas écrit à l'avance, il s'écrit au fur et à mesure par les innovateurs. C'est la société qui adopte ou pas les innovations en fonction de motivations complexes.

Aux entreprises donc de se moderniser et de créer des solutions qui, certes, répondent à des besoins métiers et exploitent l'IA, mais aussi de le faire avec responsabilité, avec les bons garde-fous pour éviter des dérives que l'on commence déjà à sentir, que ce soit au niveau du respect de la vie privée ou du simple besoin de relations humaines que nous pouvons toujours ressentir. Réduire à outrance les relations humaines sous couvert d'efficacité économique n'est pas ce à quoi l'Homme aspire naturellement.

Médias spécialisés

L'intelligence artificielle est devenue un thème couramment couvert par l'ensemble de la presse technologique, scientifique et économique généraliste. Il n'existe pas beaucoup de médias spécialisés sur l'IA en dehors de revues pointues destinées aux chercheurs.

En voici quelques exemples.

ActuIA : un site de news en français sur l'IA, www.actuia.com/acteur/list

A.I. Magazine : un site d'information en anglais sur l'IA. ai-magazine.com.

AI Magazine : un magazine US sur la recherche en IA édité par l'Association for the Advancement of Artificial Intelligence. www.aaai.org/Magazine/magazine.php.

AI Playbook : un site du fonds d'investissement Andreesseen Horowitz qui défriche le champ d'application de l'IA dans les entreprises. aiplaybook.a16z.com/

Chatbot Magazine : un site US sur les chatbot. chatbotsmagazine.com/.

In Principio, un site de vulgarisation sur l'IA couplé à un blog d'actualité, www.inprincipio.xyz.

Journal of Intelligence Artificial Research : qui comment son nom l'indique couvre l'actualité de la recherche en IA. www.jair.org/.

Journal of Machine Learning Research, www.jmlr.org.

Mais où va le web : qui commente l'actualité du numérique, dont celle de l'IA, avec une vision critique et caustique. maisouvaleweb.fr/.

Nanalyze est un bon site web faisant le tour de l'actualité autour des startups de l'IA. www.nanalyze.com.

Off the convex : un bon blog sur le deep learning, http://www.offconvex.org

Singularity Hub : magazine de l'actualité scientifique teinté par les technologies dites exponentielle, singularityhub.com/

Voicebot : un autre magazine US sur les chatbots. www.voicebot.ai/.

Arxiv : le site de publication de papiers de chercheurs coordonné par Cornell University, avec sa version dédiée à l'IA comprenant plus de 57 000 documents, www.arxiv-sanity.com/, dont la curation est assurée par Andrej Karpathy.

Dictionnaire anglais/français de l'IA

Anglais	Français	Commentaire
Back propagation	Rétro-propagation	Technique d'apprentissage de réseau de neurones
Capsule networks	Réseaux à capsules	Créés par Geoff Hinton fin 2017.
Convolutional networks	Réseaux de convolution, réseaux convolutionnels	Type de réseau de neurones pour le deep learning.
Deep learning	Apprentissage profond	Machine learning avec réseaux de neurones à grand nombre de couches.
Differential privacy	Confidentialité différentielle	Méthode d'anonymisation des données et logiciels, dont ceux de l'IA.
Feature map	*(peu traduit)*	Utilisée dans les convnets
Feed forward neuron network	Réseau de neurones avec réglage en avant	Intraduisible correctement !
Filters	Filtres	Utilisée dans les convnets
Generative Adversarial Network	Réseaux antagonistes génératifs (*appellation peu utilisée*)	Utilisé pour créer des contenus à partir d'autres contenus.
GDPR : General Data Protection Regulation	RGPD : Règlement Général de Protection des Données	Nouvelle régulation européenne de protection des données personnelles applicable depuis mai 2018.
Homomorphic cryptography	Chiffrement homomorphe	Utilisable pour faire gérer des données chiffrées dans le machine learning.
k-means	k-moyennes	Méthode de regroupement des données ou clustering dans le machine learning.
Machine learning	Apprentissage automatique	
Neuromorphic chipsets	Composants neuromorphiques	Processeurs spécialisés pour les réseaux de neurones.
Neural networks	Réseaux de neurones	Réseaux de neurones artificiels utilisés dans le machine learning et le deep learning.
Optical Characters Recognition	Reconnaissance de caractères	
Principal Components Analysis	Analyse en composantes principales	Une technique utilisée dans le machine learning.

Quantum computing	Informatique quantique	A des applications dans le machine learning.
Recurrent neuronal networks	Réseaux de neurones récurrents	Type de réseau de neurones pour le deep learning.
Reservoir Computing		Technique de réseau de neurone à base d'unités récurrentes.
Stockastic Gradient Descent	Descente stockastique de gradient	Utilisée dans la back-propagation
Shallow networks	Réseaux à faible profondeur	Utilisé dans le machine learning.
Sparse	Parcimonieux	Type de réseaux de neurones.
Spiking neurons	Neurones à impulsions, neurones impulsionnels	Utilisés notamment dans le traitement du langage et dans certains processeurs neuromorphiques.
Stacked autoencoders	Autoencodeurs empilés	Réseaux de neurones générant des contenus à partir de descripteurs.
SVM : support vector machines	Machine à vecteurs de support, séparateurs à vastes marges	Technique de segmentation du machine learning.
Tree search	Arbres de décision	Technique de machine learning pour la classification ou la régression.
Transfer learning	Apprentissage par transert	Apprentissage d'un réseau de neurones incrémental ou appliquant la reconnaissance d'un domaine d'application à un domaine voisin.
Uncanny valley	Vallée de l'étrange	Phénomène se manifestant lorsque l'on est mal à l'aise face à un robot humanoïde trop réaliste.
Variational Auto Encoder		Utilisés dans les réseaux de neurones génératifs.

Glossaire

AGI : Artificial General Intelligence, IA de niveau équivalent à celle de l'Homme. Tout du moins dans la capacité de raisonnement. C'est un Graal pour certains chercheurs et singularistes qui croient qu'il suffira d'attendre les effets de la loi de Moore pour y arriver. Les protagonistes de l'AGI oublient presque systématiquement d'indiquer à quoi celle-ci servirait et le genre de problème que l'on pourrait lui soumettre.

Alexa : service en ligne d'agent conversationnel d'Amazon, fonctionnant par reconnaissance vocale et intégré dans ses objets connectés de la gamme Echo et des objets connectés tierce-partie créés par les partenaires d'Amazon.

Algorithmes génétiques : algorithmes s'améliorant d'eux-mêmes par un processus d'évolution voisin de celui du vivant, avec des techniques de croisements et de sélection darwinienne.

ANI : Artificial Narrow Intelligence, IA utilisée dans un champ précis de résolution de problèmes. C'est l'état de l'art actuel.

ASI : Artificiel Super Intelligence, IA de niveau supérieure à celle de l'Homme. Elle serait la conséquence immédiate de l'AGI qui se reproduirait et de démultiplierait d'elle-même.

ASIC : curcuits intégrés intégrant des portes logiques gravées en dur. Les chipsets de mobiles et les microprocesseurs sont des ASIC. Ils présentent l'avantage de consommer moins d'énergie et d'être plus rapides que les FPGA mais ne sont intéressants économiquement que s'ils sont produits en grand volume. Technique utilisée par Google pour ses processeurs neuromorphiques TPU.

Back propagation : rétro-propagation, technique d'entraînement de réseau de neurones consistant à comparer le résultat du réseau sur un objet type avec la bonne classe de l'objet et de rétropropager l'erreur en remontant dans le réseau de neurones. Cela utilise des gradients, des fonctions de couts et plein de concepts divers et variés. Cette rétropropagation est réalisée pour tous les objets de la base d'entraînement. C'est un traitement très couteux en ressources machines. Il est possible de l'automatiser pour le paralléliser sur des architectures multi-cœurs ou multi-processeurs. Il est encore plus efficace dans les processeurs neuromorphiques.

Bayésien : technique d'IA s'appuyant sur des modèles probabilistes et statistiques.

BRMS : Business Rules Management Systems, les logiciels de gestion de règles permettant de créer des systèmes experts.

CNN : Convolutional Neuronal Networks, ou réseaux de neurones convolutionnels.

Connexionnisme : méthode et techniques de l'IA mettant en œuvre une modélisation à bas niveau à base de réseaux de neurones artificiels.

ConvNet : Convolutional Neuronal Networks, ou réseaux de neurones convolutionnels.

Cortana : agent conversationnel de Microsoft.

DARPA : agence américaine de financement de la R&D pour le Pentagone. L'un des plus grands financeurs de projets de R&D dans l'IA au monde. Elle finance sur challenges et appels à projet des projets qui sont réalisés par des laboratoires privés et publics, des startups, PME et grandes entreprises.

DBN : machines restrictives de Boltzmann, des réseaux de neurones datant de 1986 utilisant une seule couche de neurones source et cible et sans connexions entre les neurones d'une même couche. C'est le modèle le plus simple de réseau de neurones qui est ensuite exploité dans d'autres assemblages, comme les Deep Belief Networks (DBN) créés en 2006.

Decision Management Systems : concept marketing de système d'optimisation des décisions des entreprises qui englobe les moteurs de règles pour codifier les connaissances humaines et pratiques de l'entreprise, des modèles prédictifs qui utilisent le machine learning pour recommander les actions, et des outils d'analytics de reporting.

Deep Blue : nom de l'ordinateur qui a gagné aux échecs contre Gary Kasparov en 2007. Il s'agissait en fait d'un modèle avancé, dénommé initialement Deeper Blue.

Deep learning (apprentissage profond) : extension du machine learning intégrant des fonctions d'apprentissage supervisé et d'auto-apprentissage s'appuyant sur des modèles de représentation de données complexes et multi-dimensionnels.

Deep Mind : filiale de Google acquise au Royaume-Uni en 2014. Est à l'origine de la victoire contre le champion mondial de Go début 2016.

Feature map : composante des réseaux de neurones convolutionnels. Il s'agit d'une matrice qui contient des valeurs décrivant la pondération de l'apparition d'un filtre dans une image d'origine. Un filtre contient une forme donnée. Si celle forme est détectée, cela va donner un 1, si elle n'est pas détectée, cela donne un 0. Et toute une gradation entre zéro et un pour les valeurs intermédiaires.

Filtre : utilisé dans un réseau de neurones convolutionnel, sert à identifier des formes avec des niveaux d'abstration plus ou moins élevés. Ils sont d'abord initialisés de manière aléatoire puis ajustés progressivement par entraînement du réseau de neurones avec rétro-propagation des erreurs.

Force brute : technique de résolution de problème utilisant surtout la puissance des machines et des algorithmes traditionnels, quelle que soit leur efficacité. Souvent associée à des algorithmes dits exponentiels, dont le temps de calcul évolue de manière exponentielle avec la taille du problème à traiter.

FPGA : circuits intégrés intégrant des portes logiques qui sont définies par programmation. Ils sont adaptés à la fabrication de petites séries et au prototypage. Ils consomment plus d'énergie et sont plus lents que les ASIC.

GAN : Generative Adversarial Networks, technique de réseaux de neurones convolutionnels inversés qui génèrent des contenus à partir d'autres contenus ou d'informations élémentaires. Voir la partie qui en décrit des usages.

GOFAI : « Good Old-Fashioned Artificial Intelligence » qui dénomme les méthodes d'IA s'appuyant sur les méthodes symboliques comme dans les systèmes experts, en vogue jusque dans les années 1980.

GRU : Gated Recurrent Units, technique introduite dans les réseaux de neurones récurrents en 2014 qui simplifie les traitements par rapport au LSTM.

Hivers de l'IA : périodes de creu et de désavoeu dans l'histoire de l'IA. Le premier hiver date de la fin des années 1970 et le second de celle des années 1980 et début 1990.

IA intégrative : technique de création de solutions d'IA associant plusieurs techniques différentes (agents, moteurs de règles, réseaux neuronaux, machine learning, deep learning, bayésien, …).

Kill switch : métaphore du bouton d'arrêt d'urgence d'un ordinateur doué d'IA de niveau AGI ou ASI au cas où celui-ci ne serait plus sous contrôle.

LISP : langage de programmation d'IA utilisé dans les années 80 et 90 et notamment dans la création de systèmes experts.

Logique floue : technique d'IA créé par Lofti Zadeh dans les années 1960 et représentant l'information non pas sous forme binaire mais sous forme floue comprise entre 0 et 1. Elle est parfois utilisée dans les moteurs de règles de systèmes experts.

LSTM : long short term memory, modèle de réseaux de neurones récurrents qui intègrent bien le contexte dans lequel les éléments apparaissent de manière séquentielle. Aussi appelés réseaux de neurones à mémoire. Ils servent en particulier à interpréter le langage et à faire de la traduction automatique.

Machine learning (apprentissage automatique) : technique d'IA permettant de résoudre des problèmes de perception de l'environnement (visuel, audio, …) de manière plus efficace qu'avec les algorithmes procéduraux traditionnels. Elle s'appuie souvent sur l'usage de réseaux de neurones artificiels.

Markov, modèle de : méthode d'IA s'appuyant sur des méthodes probabilistes.

Memristors : composants électroniques à mémoire persistence. On pourrait créer avec cette technique des processeurs neuromorphiques se

raprochent le plus des neurones, avec une mémoire proche des unités de traitement permettant d'aller bien plus vite dans l'entraînement de réseaux de neurones.

Moteurs de règles : solutions techniques permettant de mettre en œuvre des systèmes experts et exploitant des bases de prédicats (règles).

Neuromorphique : se dit des processeurs neuromorphiques qui présentent la particularité d'intégrer dans leur conception des modules de calcul qui collent avec les besoins des principaux réseaux de neurones et en particulier les réseaux de neurones convolutionnels. En pratique, ils comprennent des multiplicateurs de matrices et des matrices de synapses connectant des vecteurs de neurones, plus de la mémoire locale rapide.

Pooling : technique de réduction de résolution des feature maps dans les réseaux de neurones convolutionnels. Ils permettent de réduire les temps d'entraînement et de traitements dans ces réseaux.

Rapport Lighthill : rapport anglais ayant conduit au premier hiver de l'IA en 1973 après avoir constaté les progrès trop lents de l'IA faisant suite à des promesses trop ambitieuses.

Réseaux de neurones : technique d'IA visant à simuler le fonctionnement des cellules neuronales pour reproduire le fonctionnement du cerveau humain. Est surtout utilisée dans la reconnaissance de la parole et des images. Peut-être simulé en logiciel ou avec des circuits électroniques spécialisés.

RNN : Recurrent Neuronal Networks ou réseaux de neurones récurrents. Ce sont des réseaux de neurones adaptés à l'analyse de signaux temporels comme la voix, du texte, un électro-cardiogramme ou le bruit d'une machine.

Sciences cognitives : disciplines scientifiques dédiées à la description, l'explication et la simulation des mécanismes de la pensée humaine, animale ou artificielle. Les progrès dans ces domaines permettent d'améliorer les techniques utilisées dans l'IA.

Seq2seq : sequence to sequence, technique utilisée dans le traitement du langage dans les réseaux de neurones LSTM.

SGD : stockastic gradient descent, technique utilisée dans les réseaux de neurones pour déterminer le poids optimal des synapes.

Singularité de l'IA : moment symbolique où l'IA dépassera le niveau d'intelligence humaine. Mais est-ce que cela sera un moment précis ou un continuum ?

SVM : support vector machines, technique de segmentation utilisée dans le machine learning.

Symbolisme : méthodes et techniques de l'IA visant à représenter l'information et la savoir par des concepts organisés hiérarchiquement et par relations fonctionnelles et à haut niveau.

Synapses : liaisons entre neurones au niveau de la liaison entre axones et dendrites.

Synaptique : autre appellation des processeurs neuromorphiques.

Systèmes experts : systèmes d'IA s'appuyant sur la modélisation du savoir à haut niveau avec des logiques de prédicat (si ceci alors cela, ceci est dans cela, …) et des moteurs de règles.

TPU : Tensor Processor Unit, les processeurs neuromorphiques de Google, utilisés dans leurs data centers et aussi par DeepMind pour AlphaGo.

Transhumanisme : courant de pensée ambitionnant de fusionner l'homme et la machine pour lui permettre de dépasser ses capacités intellectuelles et d'atteindre l'immortalité.

TrueNorth : processeurs neuromorphique d'IBM.

Vie artificielle : simulation de la vie à un niveau d'abstraction arbitraire, via des logiciels.

VUI : Vocal User Interface, l'interface vocale d'un agent conversationnel audio. Cela comprend l'ensemble des interactions avec l'utilisateur et leur qualité.

IBM Watson : ordinateur d'IBM ayant gagné au jeu Jeopardy en 2011. C'est maintenant une lateforme logicielle d'IA appliquée à différents métiers et besoins. (chatbot, reconnaissance d'images, etc) qui sont notamment disponibles en cloud.

Historique des révisions du document

Version	Date	Modifications
1.0 (362 pages)	19 octobre 2017	Première version publiée sur http://www.oezratty.net.
1.01	24 octobre 2017	Ajout de Neuron Data dans l'historique des systèmes experts. Mise à jour du tableau de startups marketing de l'IA provenant de Fred Cavazza.
1.02	30 octobre 2017	Corrections orthographiques diverses.
1.03	7 novembre 2017	Ajout d'In Principio dans les médias spécialisés dans l'IA.
1.04	16 novembre 2017	Remplacement de Caffee par Caffe2 dans le tableau des outils de développement. Ajout de PyTorch.
2.0 (520 pages)	15 novembre 2018	Nouvelle édition 2018 entièrement refondue et réactualisée avec 13 mois d'actualités de l'IA.
2.1 (522 pages)	16 novembre 2018	Corrections au sujet des parts de marché Amazon Echo / Google Home, au sujet d'Antvoice, et de Search'XPR. Ajout d'Ezako, du projet de trains autonomes de la SNCF, Geo4cast et Score4Biz.
2.2	17 novembre 2018	Un peu de spellcheck et modifications au sujet de Scortex et de RefundMyTicket.
2.3	20 novembre 2018	Corrections au sujet de SANEF et Vinci.
2.4	22 novembre 2018	Compléments au sujet de Dhatim et de l'expertise comptable. Ajout d'un lien sur l'histoire des IA connexionniste et symbolique.
2.5	23 novembre 2018	Ajout de Habana et SambaNova dans les chipsets serveurs.
2.6	3 décembre 2018	Corrections sur « case law » vs « civil law », et sur Deeper Blue / Deep Blue.

Vous êtes lecteur, chercheur, expert, fournisseur et avez détecté des erreurs ou graves oublis dans ce document ? Il y en a sûrement ! N'hésitez alors pas à me contacter (olivier@oezratty.net) pour me les signaler. J'effectuerai alors des mises à jour de ce rapport tout en mettant à jour la chronologie dans le tableau *ci-dessus*.

Ce document est téléchargeable à partir de https://www.oezratty.net/wordpress/2018/usages-intelligence-artificielle-2018.