Les usages de l'intelligence artificielle

Octobre 2017

Olivier Ezratty

A propos de l'auteur

Olivier Ezratty	olivier (at) 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 business plans, 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 et de conférences ou formations dans différents secteurs tels que la **télévision** (TF1, RTS-SSR, SES Astra, TDF, Euro Media Group, Netgem), les **télécoms** (Bouygues Télécom, Orange, SFR, Alcatel-Lucent), les **produits grand public** (LG Electronics, groupe Seb, L'Oréal, Alt Group), la **finance et l'assurance** (BPCE, Crédit Agricole, Crédit Mutuel-CIC, Société Générale, Natixis, Groupama). Ces missions couvrent l'assistance à la création de roadmap produit, l'analyse de positionnement et de la concurrence, la définition technologique et marketing de stratégies d'écosystèmes et « d'innovation ouverte », l'assistance à la réalisation de business plans, l'animation de séminaires de brainstorming, ainsi que l'intervention dans des conférences et séminaires sur les tendances du marché dans le numérique.

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

- Expert, membre et l'un des présidents du comité d'agrément de **Scientipôle Initiative (Wiplo)**, une association membre d'Initiative France qui accélère les startups franciliennes.
- Membre depuis fin 2015 du Comité de Prospective de l'**ARCEP**.
- Advisor du fonds d'investissement **INVST I/O** spécialisé dans les objets connectés.
- Expert auprès du pôle de compétitivité **Cap Digital** ainsi que de la **Caisse des Dépôts** et du **CNC**.
- 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 est *guest speaker* dans divers établissements d'enseignement supérieur tels que HEC, SciencePo, Neoma Rouen, CentraleSupelec, l'Ecole des Mines de Paris, Télécom Paristech et l'ECE où il intervient 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 l'auteur du **Rapport du CES de Las Vegas**, publié à la fin janvier de chaque année depuis 2006, et du **Guide des Startups** qui est devenu une référence en France avec plus de 200 000 téléchargements à date. Le tout étant publié sur le blog « Opinions Libres » (http://www.oezratty.net) qui traite de l'entrepreneuriat et des médias numériques. Comme photographe, il est aussi le co-auteur de l'initiative « Quelques Femmes du Numérique ! » (http://www.qfdn.net), devenue 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.

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 plate-forme .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 de l'Ecole Centrale Paris (1985).

Photo de couverture : schéma trouvé ici et modifié.

Table des matières

Objectifs et contenu

Par la force des choses, l'IA est devenue le sujet le plus en vue du numérique. Il est même difficile d'y échapper ! C'est à la fois un sujet de société, un objet politique et un outil de la compétitivité des entreprises. Elle génère son lot de questions sur le futur du travail, de la répartition des richesses, sur la souveraineté économique et même sur le devenir de l'espèce humaine.

Cette vague technologique est atypique par rapport aux précédentes. Les techniques de l'IA sont méconnues, 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[1]. Nombre de ses évangélistes la vulgarisent au point de saupoudrer leurs propos d'amalgames, d'approximations, de revue de presse non vérifiée 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 arriver à grand pas l'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.

Cela pour conséquence une saturation de l'espace médiatique par des contenus 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 aussi devenue un sujet central pour les entreprises, dans la lignée des autres grandes vagues technologiques du numérique telles que l'Internet ou la mobilité.

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, ce qui n'empêche pas de faire preuve de créativité et de résoudre de nouveaux problèmes de manière originale.

Cet opus fait suite à une compilation de neuf articles sur les Avancées de l'Intelligence Artificielle publiés entre mars et mai 2016 sur Opinions Libres. Depuis, pas mal d'eau a coulé sous les ponts. Le marché de l'IA se structure rapidement. Les briques technologiques de l'IA prennent forme et sont assemblées par les uns et les autres, éditeurs de logiciels, fournisseurs de solutions en cloud, startups et sociétés de services. Le marché des composants spécialisés dans l'IA prend aussi forme, des data centers aux objets connectés en passant par les smartphones.

---

[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 est déjà disponible et opérationnelle, sans d'ailleurs forcément s'en rendre compte eux-mêmes !

Cette édition me permet d'ailleurs de corriger un bon nombre d'erreurs techniques ou d'appréciations qui émaillaient la première édition de ce texte ainsi que des lacunes importantes pour ce qui concernait les techniques du deep learning. En appréhender le fonctionnement pour pouvoir en saisir le potentiel mais aussi les limites demande un peu d'effort. Est-ce que le deep learning est si miraculeux et universel que le marketing le laisse à penser ? Et bien non ! Son approche est essentiellement probabiliste et n'a rien d'un raisonnement formel.

Qui plus est, l'intelligence artificielle n'est pas un produit. Elle ne se présente 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 d'intelligence artificielle très 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.

Nous en sommes toujours à l'âge de pierre, avec seulement une soixantaine d'années de recul sur la question et une dizaine d'années pour ce qui est du deep learning. Mais les chercheurs font rapidement avancer le domaine 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 il existe de nombreux jeux de données en open data pour entrainer ses modèles !

L'IA est un grand tonneau des Danaïdes. 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". Et 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.

L'IA rassemble un grand nombre de spécialités. Selon certains, il faudrait un PhD en IA pour pouvoir développer une solution d'IA. C'est peut-être vrai aujourd'hui mais de nombreux outils 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.

J'ai bien conscience de l'escroquerie intellectuelle consistant à faire croire que j'ai tout compris. Rassurez-vous : ce n'est pas le cas et pas mal de domaines de l'IA m'échappent encore et ma besace est pleine d'interrogations diverses, en particulier autour des techniques de traitement du langage[4] !

---

[3] Cf 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.

Contrairement à sa première version, ce document adopte un découpage en parties qui est plus 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.

Voici la synthèse et la structure de ce document qui est organisé en huit grandes parties :

- **Grands courants de l'IA :** qu'est-ce que l'IA ? 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 ? 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 ?
- **Briques fondamentales de l'IA** : quelles sont les principales briques mathématiques et algorithmiques de l'IA ? Les progrès récents viennent-ils du logiciel, du matériel ou des données ? Qu'est-ce que le machine learning et comment le met-on en œuvre d'un point de vue pratique ? Quid du deep learning et de ses nombreuses variantes ? Et les systèmes experts, pourquoi en parle-t-on moins que pendant les années 1980 ? Comment les briques d'intelligence artificielle progressent-elles ? Quels sont les outils de création d'applications d'IA et pourquoi la majorité sont-ils open source ? Comment fait-on de la programmation en IA ? Quid de l'intelligence artificielle généralisée ? Est-ce un fantasme ? Peut-on facilement reproduire le fonctionnement du cerveau humain ?
- **IA et infrastructure informatique** : quelles sont les ressources matérielles qui font avancer l'IA ? Comment évolue l'application de la loi de Moore ? Pourquoi fait-on maintenant appel à des GPU et à des processeurs neuromorphiques pour les applications de l'IA ? Comment se distinguent-ils ? Quels sont les nouveaux acteurs de ce marché ? Pourquoi il y-a-t-il une grande différence entre l'entrainement d'une IA et son exécution ? Est-ce que l'informatique quantique aura un impact sur l'IA ? Quel est le rôle des capteurs et des objets connectés ? Comment sont gérées les ressources en cloud de l'IA ainsi que du côté des systèmes embarqués ?
- **Applications génériques de l'IA** : quelles sont les applications génériques et horizontales de l'IA, dans le traitement de l'image, du langage et de la parole, dans la robotique, dans le marketing, les RH 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 ? Pourquoi certains de ces marchés sont plus dynamiques que d'autres ? Comment les startups permettent aux entreprises d'innover dans ces différents marchés[5] ?

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

- **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 ? 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 ?
- **L'IA dans la société** : pourquoi les points de vue sur l'impact potentiel de l'IA sur les métiers et sur la société en général sont-ils si variés et contradictoires ? Comment l'IA et la robotique vont transformer les métiers dans le futur ? Est-ce un tsunami qui se prépare ? Que disent les experts sur le sujet ? 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 ?
- **L'IA et l'entreprise** : comment les entreprises peuvent-elles intégrer l'IA dans leur stratégie ? Quelles sont les bonnes pratiques ? Comment gérer les compétences ? Peut-on benchmarker l'IA ? Comment s'organiser ? Comment intégrer l'IA dans les autres dynamiques d'innovations liées au numérique ? Comment va évoluer le métier de développeur ?

Voilà le programme !

Je m'appuie en grande partie 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 ! Mais il y a évidemment probablement plein de trous dans la raquette que les esprits avertis ne manqueront pas de souligner !

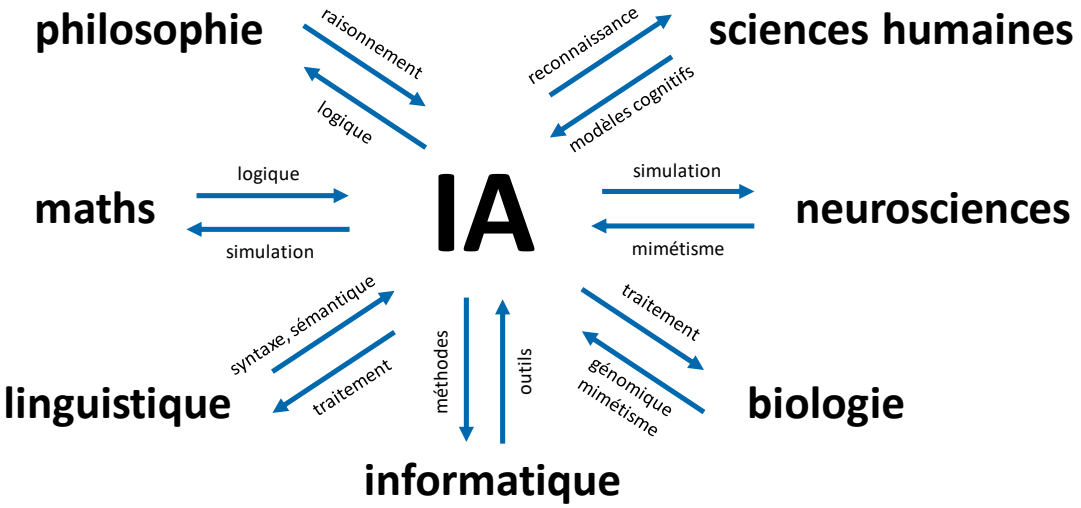
Bonne lecture !

Grands courants de l'IA

L'intelligence artificielle génère toutes sortes de fantasmes pour les uns et de craintes pour les autres. Les définitions de l'IA vont d'ailleurs bon train. La notion d'intelligence est elle-même source d'une interminable bataille sémantique[6] qui touche notamment les startups du numérique[7].

Pour certains, seul l'apprentissage profond ou deep learning est digne de faire partie de l'IA et le machine learning, pas du tout. Comme si seules les technologies un peu magiques dans leur apparence pouvaient faire partie de l'IA. Pour les spécialistes du secteur, toutes les technologies de l'histoire de l'IA en font partie. Seules leurs capacités évoluent dans le temps. Ces différents viennent aussi de ce qu'une partie importante de la valeur ajoutée des applications de l'IA est issue du volume et de la qualité des données qui les alimentent. Mais les méthodes importent tout autant et elles évoluent constamment.

L'intelligence artificielle représente un pan entier de l'informatique avec sa diversité, ses briques technologiques, ses méthodes, ses assemblages et solutions en tout genre. Elle est aussi intimement liée à d'autres sciences : les mathématiques et les statistiques qui lui servent de base théorique, les sciences humaines (sciences cognitives, psychologie, philosophie, …) 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.



[6] 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.

[7] La querelle sémantique qui atteint l'univers des startups bat son plein. 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 custom 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. Ce qui permet de faire un premier tri.

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.

Seules les briques technologiques de base sont génériques, mais utilisées par les développeurs, comme **TensorFlow**, **PyTorch** ou **Theano**.

Hauts et bas de l'IA

L'IA a pris forme conceptuellement avec le concept de calculus ratiocinator de **Leibnitz** (circa 1671), la machine et le fameux test de **Turing** que l'on ne présente plus (1950), les neurones formels de **McCulloch** et **Pitts** (1943), l'architecture de **Von Neuman** (1945) ou encore le théorème de **Shannon** (1949).

Son histoire moderne a cependant véritablement démarré au moment du **Summer Camp de Darmouth** de 1956. Il s'agissait d'une sorte de hackathon intellectuel de près de deux mois réunissant une dizaine de scientifiques.

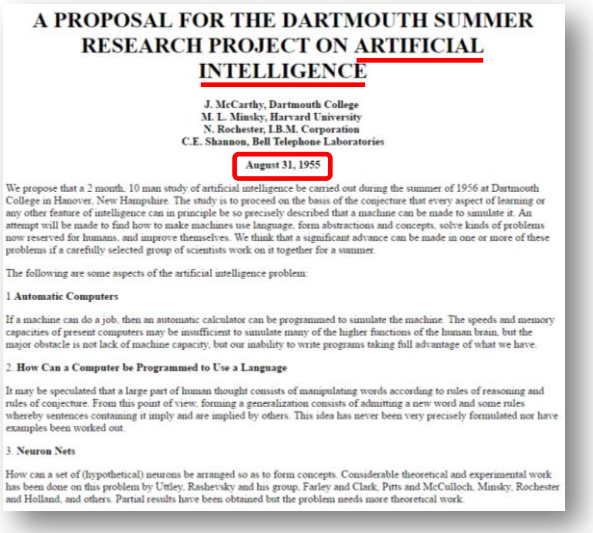
L'IA, un voyage éternel ou un aboutissement ?

L'expression « intelligence artificielle » fut proposée en 1955 par l'un des initiateurs de ce summer camp, **John McCarthy**[8]. 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 définit l'IA comme le champ académique de création de logiciels et matériels doté de certaines formes d'intelligence.

Summer Camp de Darmouth

USA New Hampshire – été 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**

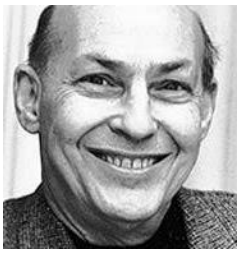
If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability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

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

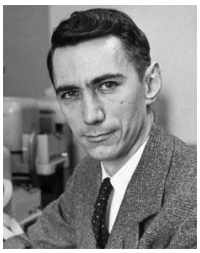
It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalization consists of admitting a new word and some rules whereby sentences containing it imply and are implied by others. 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, MIT
1927-2016



Claude Shannon, Bell
1916-2001

John McCarthy, MIT
1927-2011

Nathaniel Rochester, IBM
1919-2001

[8] Pour la petite histoire, 1955 est aussi l'année de la naissance de Steve Jobs et Bill Gates.Tout un symbole ! A l'époque, les ordinateurs étaient des plus rares et fonctionnaient avec des lampes en lieu et place des transistors qui ont fait leur apparition pendant les années 1960 dans les ordinateurs, notamment dans la série 360 d'IBM.

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.

L'appellation est à l'origine de débats épistémiologiques sans fin sur ce qu'est l'IA et sur la manière de la comparer à l'intelligence humaine. L'IA décrit aussi bien le champ du possible d'aujourd'hui dans ces domaines que la quête permanente et insatisfaite de l'intégration de l'intelligence humaine dans les machines.

L'appellation la plus appropriée serait peut-être celle d'intelligence humaine augmentée, l'IA étant principalement destinée à permettre à l'homme de faire plus de choses, 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. Dans le domaine du raisonnement automatisé, l'IA est censée apporter une rationnalité dont l'Homme ne fait pas toujours preuve. Là encore, nous somme dans l'ordre de la complémentarité.

L'IA fait partie de ce que l'on appelle aussi les sciences cognitives. Elles comprennent d'abord les sens et la capacité des ordinateurs à lire, voir et entendre, puis à structurer leur mémoire, à apprendre, à raisonner, puis à prendre des décisions ou à aider à prendre des décisions.

Le groupe de travail du summer camp de Darmouth comprenait **Marvin Minsky**, **Claude Shannon**, à l'époque au MIT, **Allan Newell** et **Herbert Simon** de Carnegie Tech, et **Arthur Samuel** et **Nathanial Rochester**, tous deux d'IBM. Le groupe voulait plancher sur la conjecture selon laquelle tous les processus de l'intelligence humaine pouvaient théoriquement être mis en œuvre par des machines. Les discussions étaient surtout conceptuelles. Ces chercheurs pensaient aboutir rapidement à un résultat probant. Plus de 60 ans après, nous y sommes encore !

L'IA est 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[9]. 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[10].

[9] 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.

[10] Source de cette énumération : cours d'intelligence artificielle d'Olivier Boisard.

Les bases conceptuelles de l'IA d'aujourd'hui datent des années 1950 !

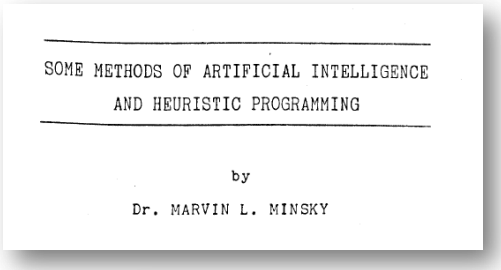
Deux ans après le summer camp de Darmouth avait lieu le **Congrès de Middlesex** (1958) au Royaume Uni avec des contributions des principaux artisans du Congrès de Darmouth, Marvin Minsky et John MacCarthy ainsi qu'**Oliver Selfridge**, lui aussi présent à Darmouth.

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.

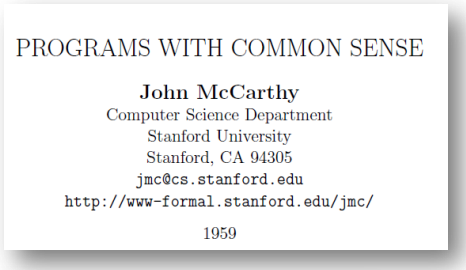
congrès de Middlesex - UK - 1958

mechanization of thought processes


SOME METHODS OF ARTIFICIAL INTELLIGENCE
AND HEURISTIC PROGRAMMING

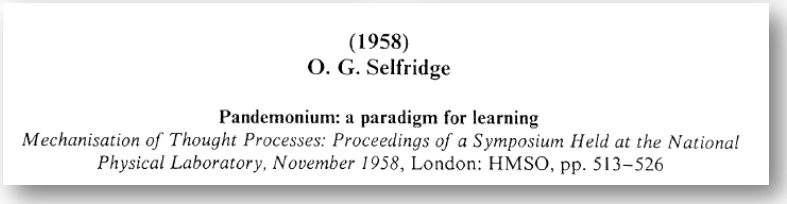
by

Dr. MARVIN L. MINSKY


PROGRAMS WITH COMMON SENSE

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1959


(1958)
O. G. Selfridge

Pandemonium: a paradigm for learning
Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory, November 1958, London: HMSO, pp. 513–526

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 de budget annuel actuellement.

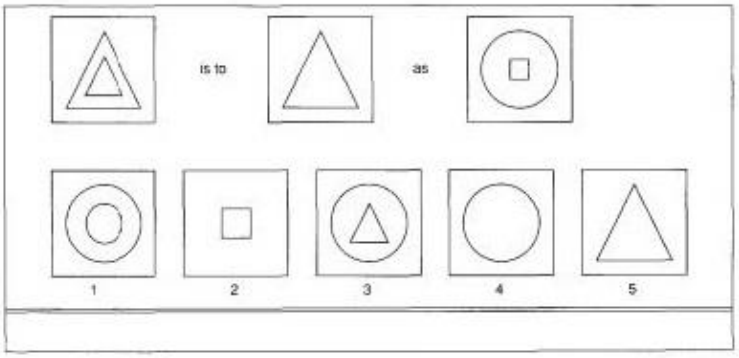
La recherche sur l'IA était financée par les deniers publics, notamment aux USA et au Royaume-Uni. Encore aujourd'hui, une très grande partie des recherches les plus avancées sur l'IA aux USA le sont par l'omniprésente DARPA ainsi que par les agences liées au renseignement comme la CIA et la NSA. Ce 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).

Les 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.

Il y eu le **Geometry Theorem Prover** d'Herbert Gelernter en 1959, un logiciel de démonstration de théorèmes de géométrie fonctionnant en chainage 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 plutôt simple. C'était plutôt prometteur.

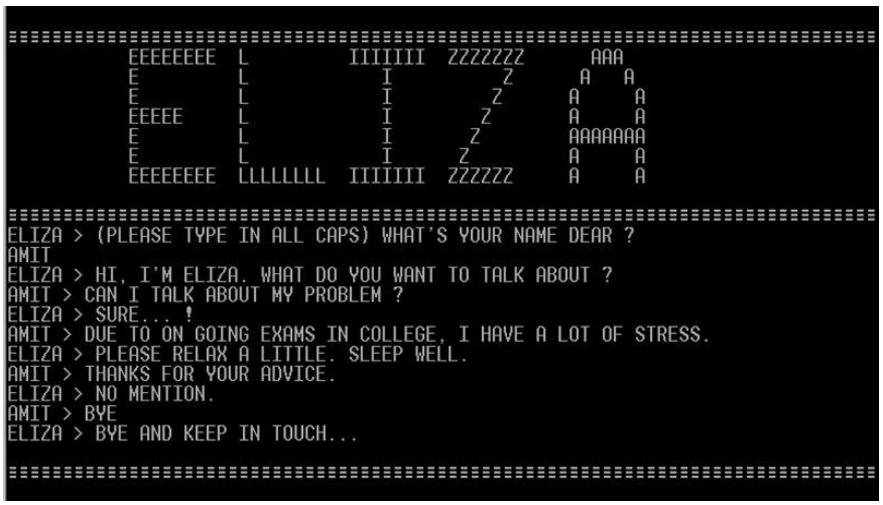


Suivirent le **General Problem Solver** d'Allen Newell et Herbert Simon en 1959, l'**Integration Problems Solver** de James Slagles en 1963, le **Geometric Analogy Problems** de Tom Evans en 1968, qui traitait les problèmes de géométrie qui sont intégrés dans les tests de quotient intellectuel (*ci-dessus*) 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.

Les premiers chatbots datent des années 1960 !

On vit aussi apparaitre 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, puis **SHRDLU**, de Terry Winograd du MIT, l'un des premiers à comprendre le langage naturel en 1968.

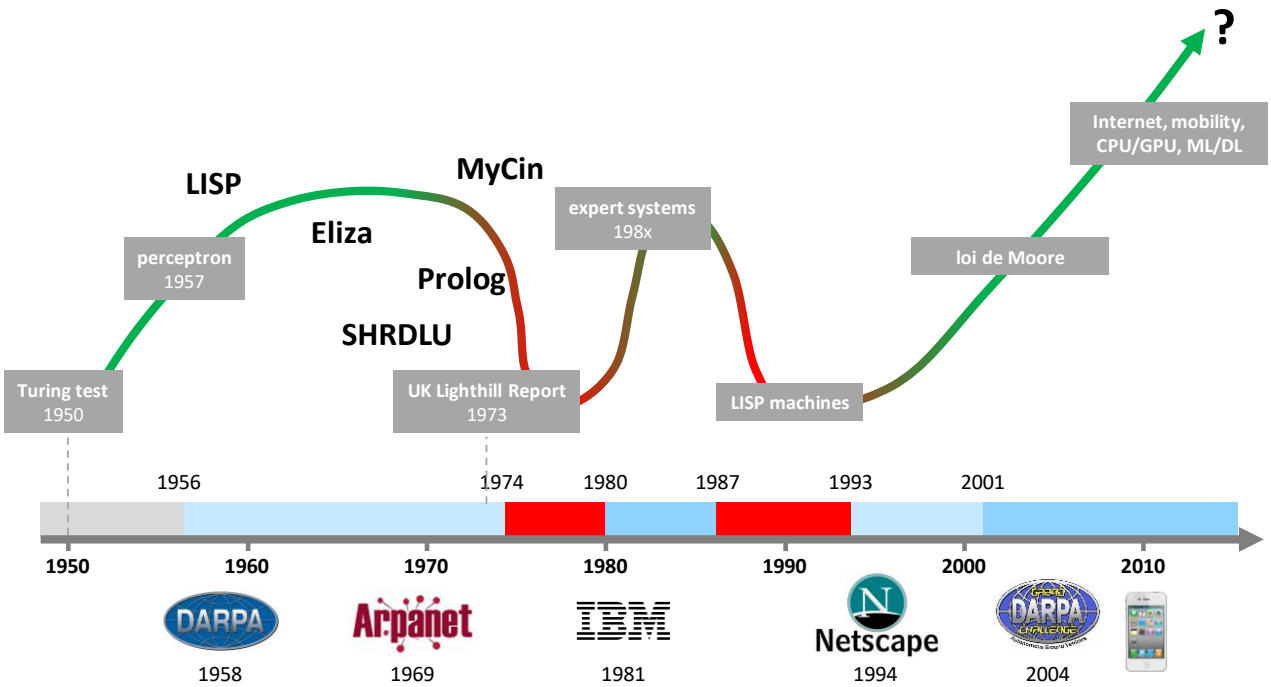


Ces premiers chatbots tenaient le coup pendant des conversations avec quelques échanges mais ne passaient pas le test de Turing. Malgré tout, ils n'ont pas à rougir vis-à-vis de nombreux chatbots contemporains.

Les surpromesses et le premier hiver de l'IA

L'IA connu son premier "hiver" avec une réduction d'une bonne part de ses budgets de recherche à partir de 1973, tant au Royaume-Uni qu'aux USA.

C'était notamment 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[11]. 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.

Cette **histoire de l'IA**[12] en fait un inventaire intéressant. **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 !

*"Within our lifetime **machines may surpass us in general intelligence**."*
Marvin Minsky 1961

*"Machines will be capable, within twenty years, of **doing any work a man can do**."*
Herbert Simon 1965

*"Within a generation ... the problem of **creating 'artificial intelligence' will substantially be solved**."*
Marvin Minsky 1967

*"In from **three to eight years** we will have a machine with the **general intelligence** of an average human being."*
Marvin Minsky 1970

"In 2019, **household robots are ubiquitous and reliable**."
Ray Kurzweil, 1999

*"Within a decade, **AIs will be replacing scientists and other thinking professions**."*
John Hall 2011

On attend toujours la démonstration par une IA évoluée du théorème d'incomplétude de **Godel** qui dit que "*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>2).

Le théorème de Fermat a été démontré au milieu des années 1990 et après des années d'efforts de plusieurs mathématiciens dont **Andrew Wiles**. Sa démonstration publiée dans les annales de mathématiques 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.

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 sans apprentissage supervisé, il y aura vraiment avoir de quoi être bluffé !

---

[11] 1973 est l'année de l'apparition du premier micro-ordinateur de l'histoire, le français Micral de François Gernel et André Truong.

[12] http://kuliah-sore-malam-unkris.ggkarir.co.id/IT/en/2185-2061/history-of-artificial-intelligence_9498_kuliah-sore-malam-unkris-ggkarir.html

Pierre de Fermat

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

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.

Herbert Simon 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 dans la pratique 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 ou 2045 et l'immortalité ou une vie de 1000 ans pour les enfants qui viennent de naitre !

Le premier hiver de l'IA a duré jusqu'en 1980. Il correspond d'ailleurs au premier âge de l'industrie de la micro-informatique, avec la création de Microsoft (1975), d'Apple II (1977), d'Oracle (1977) puis les préparatifs du lancement de l'IBM PC (1980-1981).

La première renaissance avec les systèmes experts et un 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.

Créé par les français Alain Colmerauer[13] et Philippe Roussel en 1972, le langage **Prolog** a participé à cette vague.

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.

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 des dépenses d'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 en 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.

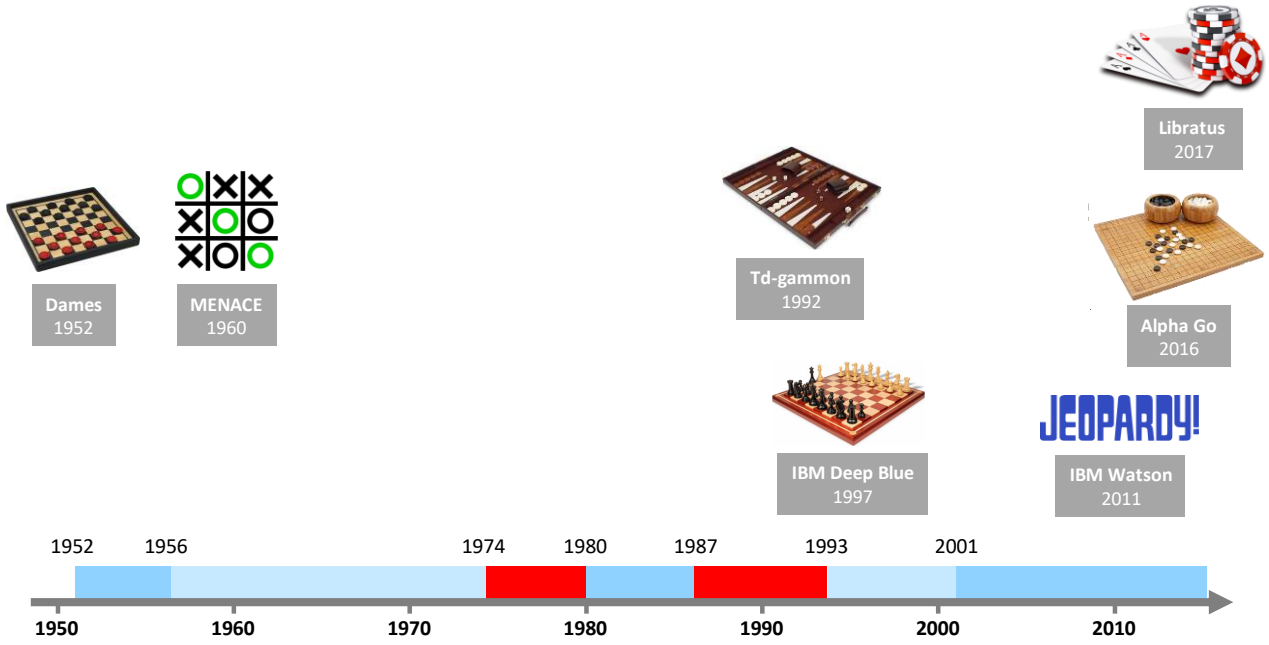
La dernière renaissance de l'IA

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 les évoquer dans la seconde partie dédiée aux techniques 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 de méthodes statistiques 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.

[13] Alain Colmerauer est décédé en mai 2017.

- L'atteinte de diverses **étapes symboliques** 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 son 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[14], 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.



- 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.

[14] Cf 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 rticle 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 !

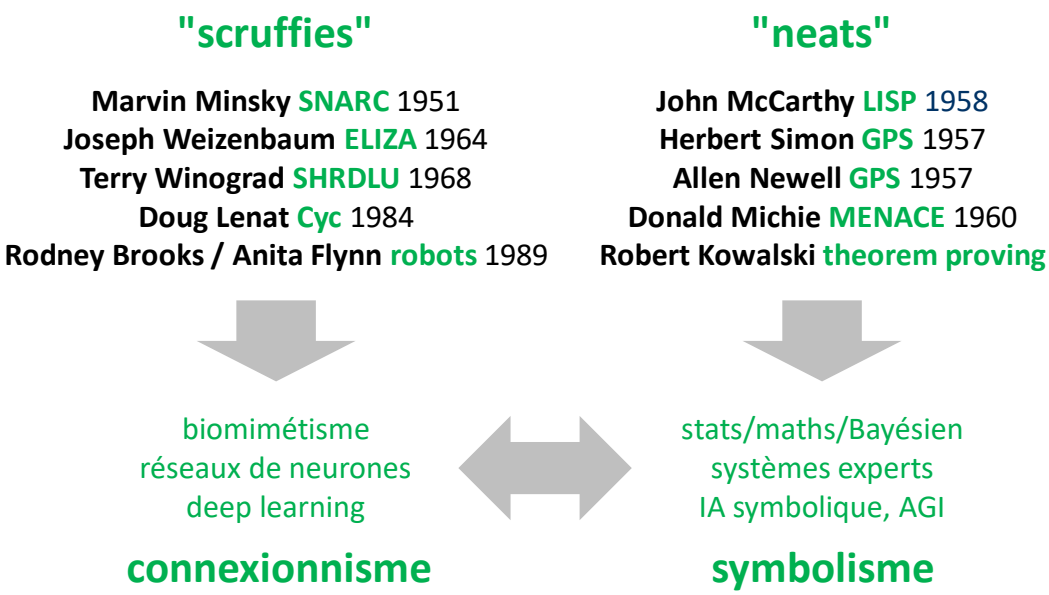
- La disponibilité de très **gros volumes de données**, via les usages de l'Internet et des mobiles, des objets connectés ou de la génomique, exploitables par différentes méthodes de machine learning et de deep learning. Le travail des chercheurs et développeurs est facilité par la publication de jeux de données ouverts (open data) pour le deep learning avec de nombreuses bases d'images et de textes disponibles pour réaliser des benchmarks. C'est le cas de la base d'ImageNet, de la base décriture manuscrite MNIST et de la base linguistique WordNet (en anglais). Ces bases sont généralement d'origine américaine. Elles proviennent le plus souvent de grandes Universités.
- La culture de l'**open source** qui domine les outils de développement de solutions d'IA. Les chercheurs publient de plus en plus des exemples de codes sources pour illustrer leurs méthodes, sur Github. Ils sont alors reproduits et vérifiés par la communauté des développeurs. Ce processus permet une diffusion rapide des nouveautés algorithmiques, particulièrement autour des réseaux de neurones et du deep learning.
- 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.
- Les **nombreuses applications commerciales** de l'IA croisant le machine learning, les objets connectés, la mobilité et le big data. Avec des attentes fortes dans le marketing, le e-commerce et la finance.

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.

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.

On a vu au départ s'opposer les partisans du connexionnisme – utilisant le principe du biomimétisme, des réseaux de neurones et de l'auto-apprentissage qui est pour l'instant surtout utilisé pour les sens artificiels – face à ceux de l'intelligence artificielle symbolique 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, …).



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. Comme si il y avait un écart entre la côté Est et la côte Ouest !

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 pré-conditionnement 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, comme nous le verrons dans le dernier article de cette série, à en découvrir 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 pendant depuis le milieu des années 2000-2010, en particulier pour la vision artificielle et la reconnaissance et le traitement du langage. La technique la plus remarquable étant celle des réseaux de neurones convolutionnels, créée par le français **Yann LeCun**.

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

Selon **Pedro Domingos**, l'auteur de « The Master algorithm », il existe en fait cinq grands courants dans l'IA en plus du symbolisme et du connexionnisme (*ci-dessus*). 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. Et dans de nombreux cas, ces approches sont combinées pour générer des solutions optimales.

Définitions et segmentations de l'intelligence artificielle

L'IA est un ensemble de techniques permettant d'imiter le comportement humain, agissant de manière rationnelle en fonction de faits, données et expériences, et capables d'atteindre un ou plusieurs objectifs donnés de manière optimale.

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.

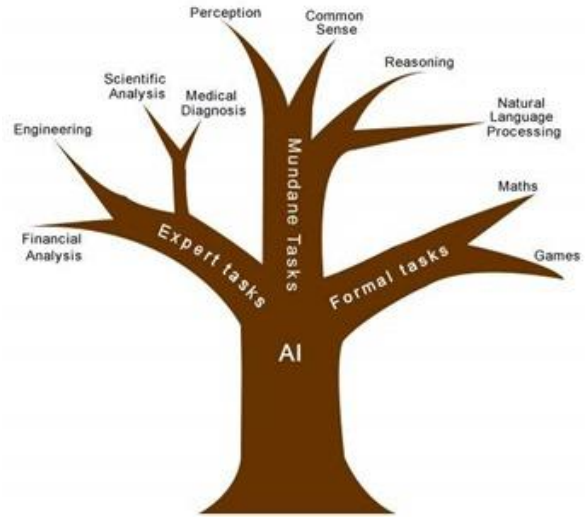
Dans mes recherches bibliographiques sur l'IA démarrant début 2016, j'ai cherché à segmenter le vaste champ de l'IA.

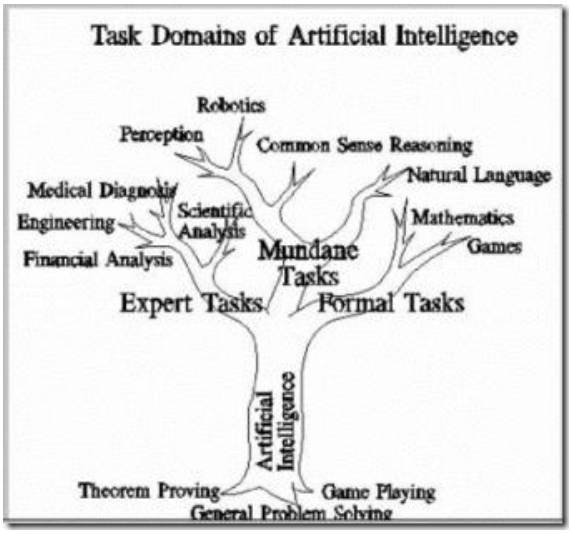
A haut niveau et scientifiquement parlant, on peut le découper en trois grands domaines :

- Le **symbolisme** qui se focalise sur la pensée abstraite et la gestion des symboles. C'est dans cette catégorie que se trouvent les systèmes experts et dans une certaine mesure, le web sémantique. 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.
- Le **connexionisme** 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 « micro » de résolution des problèmes. C'est ici que l'on peut ranger le deep learning utilisé dans la vision artificielle ou la reconnaissance de la parole.
- 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 reconnaitre, exprimer, synthétiser et

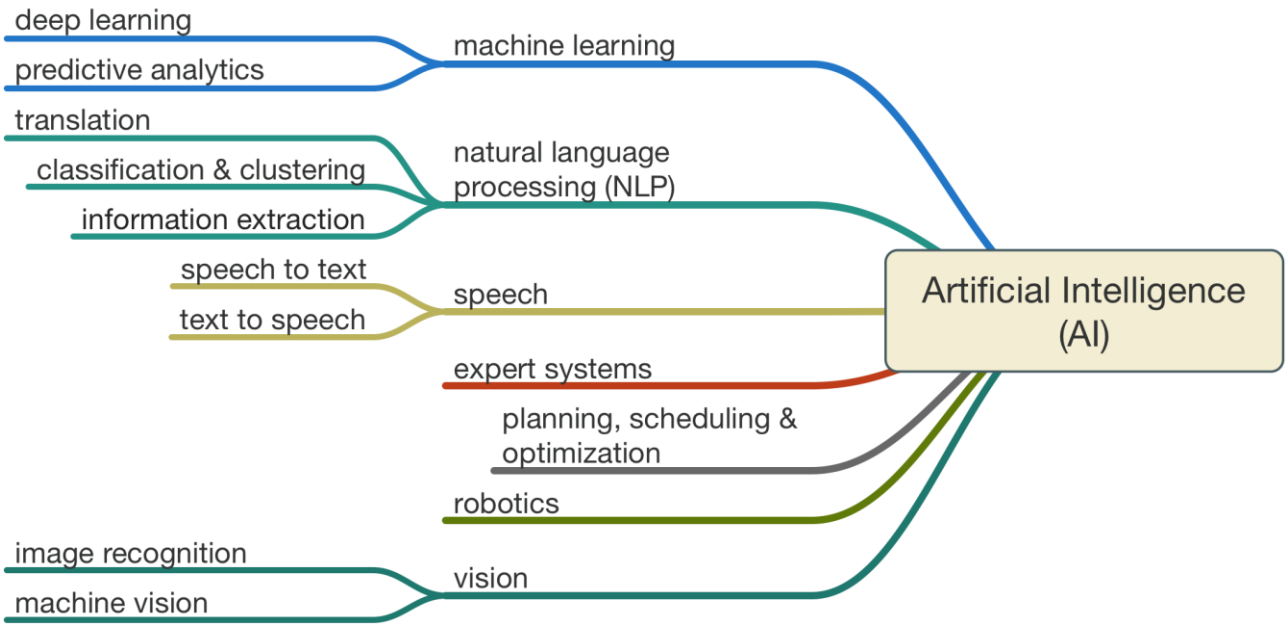
modéliser les émotions humaines. C'est une capacité qu'IBM Watson est censé apporter au robot Pepper de/ Softbank Robotics (ex Aldebaran).

Reprenant plus ou moins ce découpage, cet autre découpage 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).

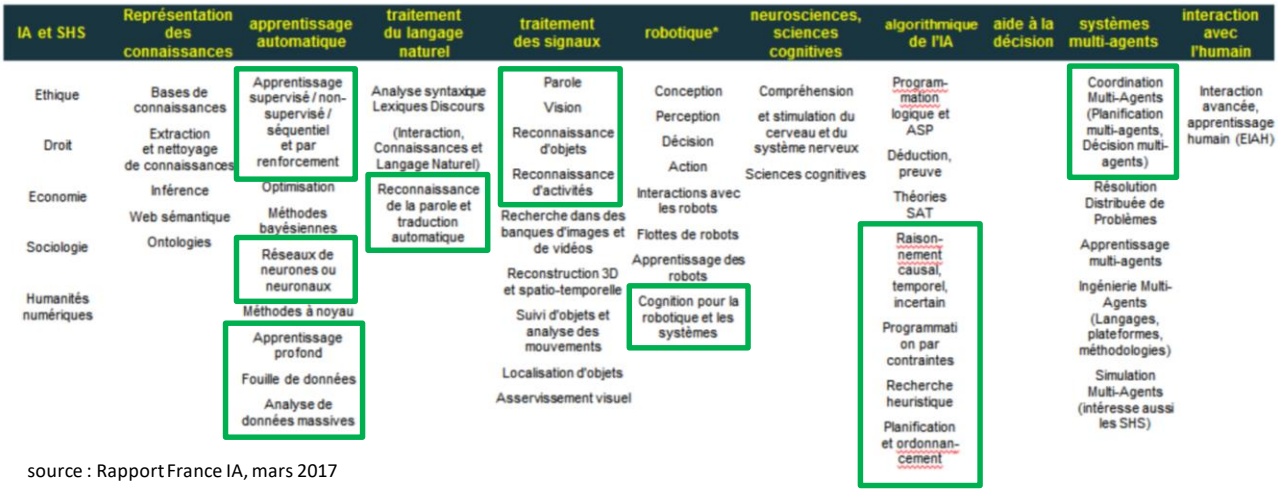




Cette autre segmentation très utilisée comprend le machine learning, le deep learning, le traitement du langage, les systèmes experts, la robotique et la vision. Il met curieusement au même niveau des outils génériques comme le machine learning et le deep learning et ses propres applications comme la 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.



Le rapport **France IA** publié en mars 2017[15] par le gouvernement propose pour sa part une segmentation plus fouillée, compilant les principaux travaux de recherche du domaine en France.

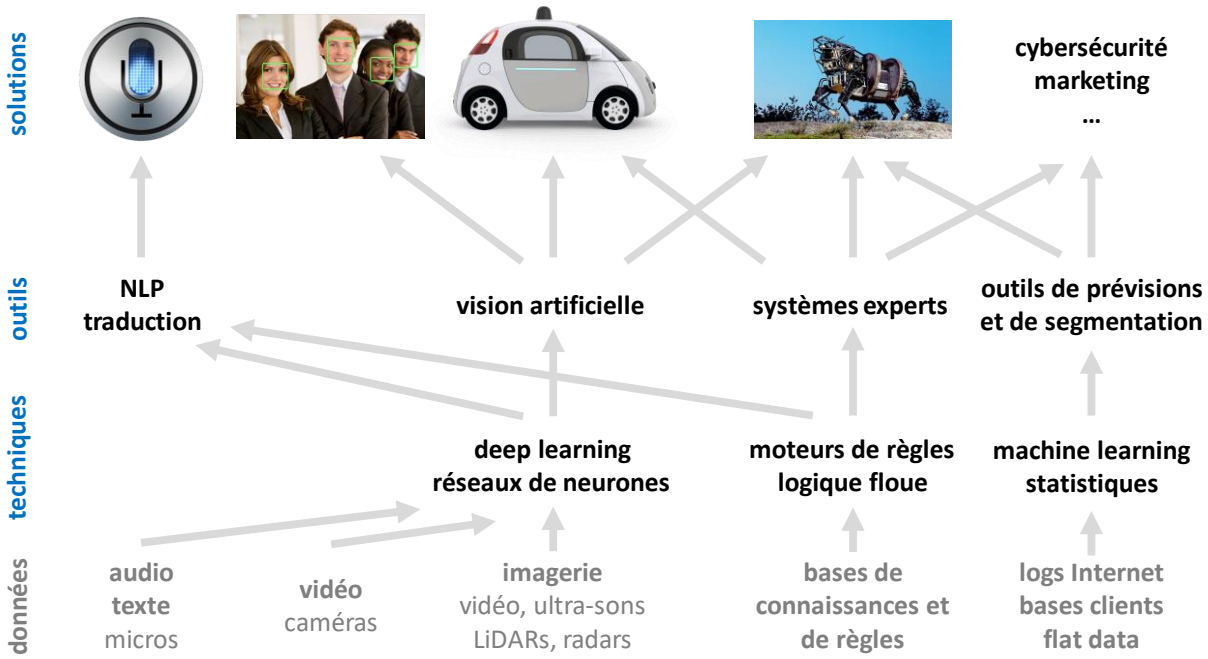


source : Rapport France IA, mars 2017

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 chatbot, les véhicule autonomes, les robots, les systèmes de recommandation, les outils de segmentation client, le marketing prédictif ou les solutions de cybersécurité.
- 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.
- 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.

[15] Ici : http://www.enseignementsup-recherche.gouv.fr/cid114739/rapport-strategie-france-i.a.-pour-le-developpement-des-technologies-d-intelligence-artificielle.html.



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[16].

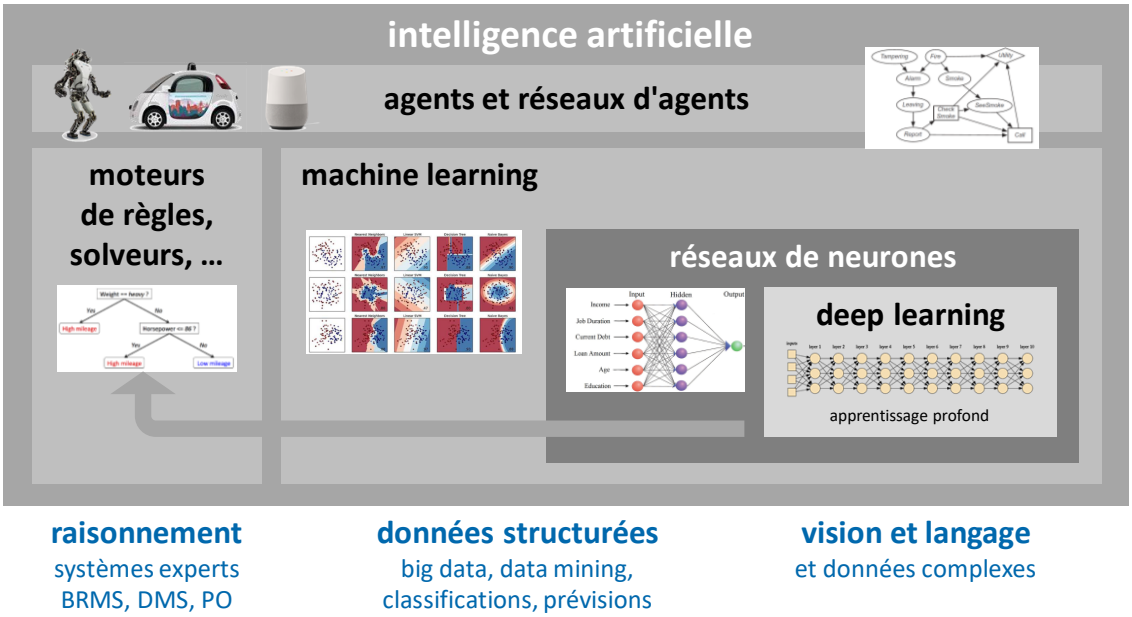
Quand une startup indique qu'elle a créé « une IA » pour faire ceci ou cela, cela signifie qu'elle a assemblé des techniques, paramétré des outils, en général assez standards, pour exploiter des données, et les a appliqués pour créer une solution. L'originalité est rarement technique, mais plutôt dans la verticalité de l'assemblage !

[16] Aymeric Poulain Maybant m'a transmis sa **thèse de doctorat** sur 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.

Briques fondamentales de l'IA

Reprenant mon schéma hiérarchique de la partie précédente, je vais maintenant partir des couches d'abstraction les plus basses (systèmes experts, réseaux neuronaux, machine learning, méthodes statistiques, …) pour remonter dans les parties suivantes dans les couches plus hautes qui font généralement appel aux couches basses, comme dans la reconnaissance de la parole ou des images dans la partie suivante.

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.



Il me semble à ce stade utile de départager quatre grandes briques couramment utilisées dans les solutions d'IA :

- Les **moteurs de règles** qui permettent de construire des systèmes experts à base de règles formelles avec des logiques dites symboliques. Cette technique fait beaucoup moins parler d'elle car elle a connu des hauts et des bas et le deep learning a submergé l'espace médiatique de l'IA. Mais elle reste fondamentale pour un grand nombre de systèmes. 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.
- Le **machine learning** qui permet de faire des prévisions, de la classification et de la segmentation automatiques en exploitant des données en général multidimentionnelles, comme une base de données clients. Le machine learning relève d'une approche probabiliste. Les outils du machine learning servent à exploiter le « big data ». Le machine learning peut s'appuyer sur des réseaux de neurones simples pour les tâches complexes portant sur des données multidimentionnelles.

- Le **deep learning** ou apprentissage profond, qui permet de gérer un niveau d'abstraction plus élevé que le machine learning afin de reconnaitre des objets complexes comme les images, l'écriture manuscrite ou la parole. Le deep learning s'appuie sur 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[17]. Le deep learning permet aussi de générer des contenus ou d'améliorer des contenus existants, comme pour coloriser automatiquement des images en noir et blanc.
- Les **réseaux d'agents**, un domaine méconnu qui couvre la science de l'orchestration des briques techniques de l'IA pour créer des solutions. 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 logiciels d'assemblages de briques logicielles.

Dans le schéma *ci-dessus*, je relie le deep learning avec les systèmes experts. Pourquoi donc ? Parce qu'historiquement, il était difficile d'alimenter à la main les systèmes experts avec des règles. Le deep learning permet d'examiner de gros volumes de textes et données et d'en déduire des règles qui, à leur tour, peuvent alimenter des moteurs de règles et des systèmes experts.

Force brute et arbres de décision

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. Cela peut fonctionner si c'est à la portée de la puissance de calcul des machines.

Ces mécanismes rudimentaires sont optimisés avec des algorithmes d'élagage qui évacuent les "branches mortes" de la combinatoire ne pouvant aboutir à aucune solution. C'est d'ailleurs plus facile à réaliser aux échecs qu'au jeu de Go car la combinatoire du premier est plus faible que celle du second !

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

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

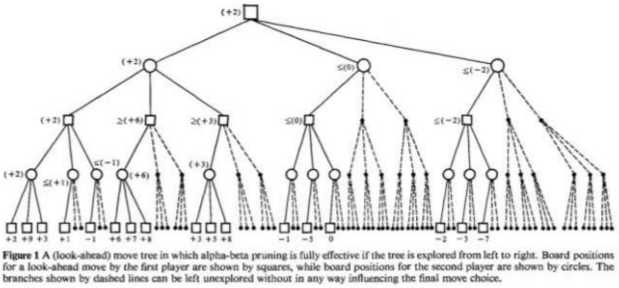


Figure 1 A (look-ahead) move tree in which alpha-beta pruning is fully effective if the tree is explored from left to right. Board positions for a look-ahead move by the first player are shown by squares, while board positions for the second player are shown by circles. The branches shown by dashed lines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

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

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.

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. 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[18].

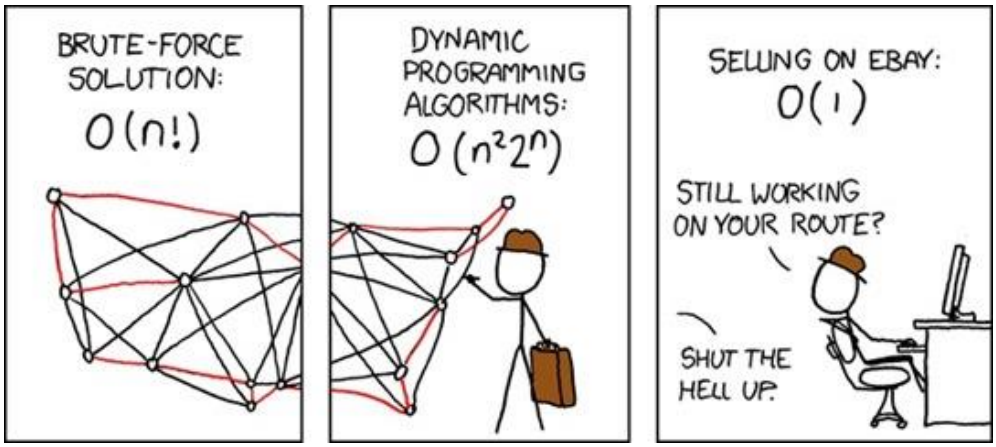
8x8

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 est utilisée dans de nombreux domaines comme dans les moteurs de recherche ou la découverte de mots de passe. 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.



(**source de l'image**)

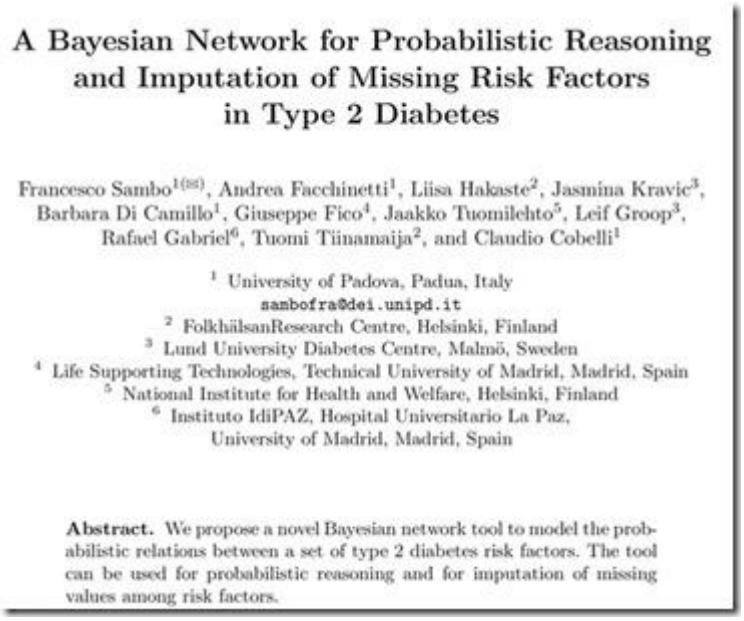
[18] En octobre 2017, une nouvelle version d'AlphaGo dite Zero gagnait contre la version de début 2017 sans avoir à être entrainée avec des parties jouées par des humains et avec comme seule information de départ les règle du jeu de Go et la position initiale des jetons. La méthode relève toujours de 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. Cf Intelligence artificielle : toujours plus puissant, AlphaGo apprend désormais sans données humaines et l'article technique de 42 pages qui documente la prouesse et AlphaGo Zero: Learning from scratch, de DeepMind, qui vulgarise la performance.

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*).

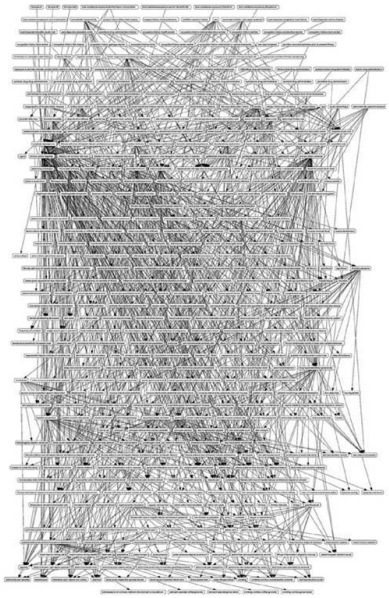


A Bayesian Network for Probabilistic Reasoning and Imputation of Missing Risk Factors in Type 2 Diabetes

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.



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.

Le cerveau met d'ailleurs en œuvre une logique bayésienne pour ses propres prises de décision, notamment motrices, les centres associés étant d'ailleurs situés dans le cervelet tandis que dans le cortex cérébral gère la mémoire et les actions explicites[19].

[19] Source : **Stanislas Dehaene**.

Dans la pratique, les méthodes statistiques se sont fondues avec le temps dans les techniques du machine learning et de deep learning. Ces dernières reposent en effet tous sur des modèles probabilistes pour identifier des objets ou prédire le futur. Seule l'IA symbolique qui s'appuie sur de la logique formelle ne relève pas des probabilités.

Systèmes experts

Les systèmes experts font partie du vaste champ de l'IA symbolique appliquant de la logique formelle.

Elle s'oppose 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.

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[20].

Les 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.

En 1958, **John McCarthy** créait le langage de programmation LISP adapté 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**[21].

En 1959, **Herbert Gelernter** créait le Geometry Theorem Prover[22], 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 créé dans les années 1970 capable de déterminer des réactions chimiques de synthèse de molécules organiques.

---

[20] Cf Intelligence Artificielle Symbolique de Guillaume Piolle, 2015.

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

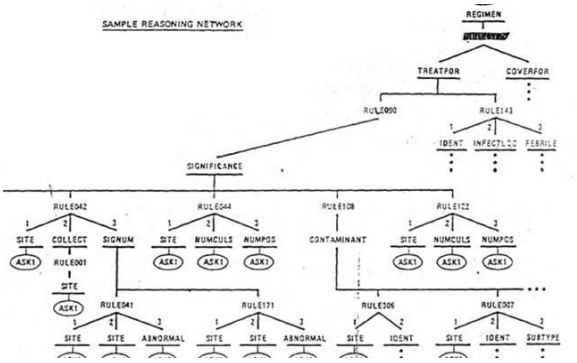
[22] Cf https://pdfs.semanticscholar.org/2edc/8083073837564306943aab77d6dcc19d0cdc.pdf.

Dans les années 1970, **Robert Kowalski** de l'Université d'Edinbourg fit avancer les travaux dans la programmation logique. Puis les français **Alain Colmerauer** et **Philippe Roussel** créèrent le langage de programmation **Prolog** qui connut ses heures de gloire dans les années 1980.

Les 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 symptomes (*ci-dessous*) avec une base de 450 règles.

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.




```
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
```

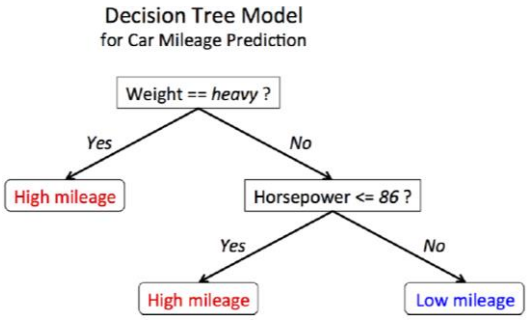
Les moteurs de règles et les solveurs sont couramment employés dans les systèmes experts depuis les années 1980[23]. Et ils ont connu de nombreux progrès (*ci-dessous*) malgré l'hiver de l'IA de la fin des années 1980 et débuts 1990. C'était surtout un hiver des systèmes experts et du LISP !

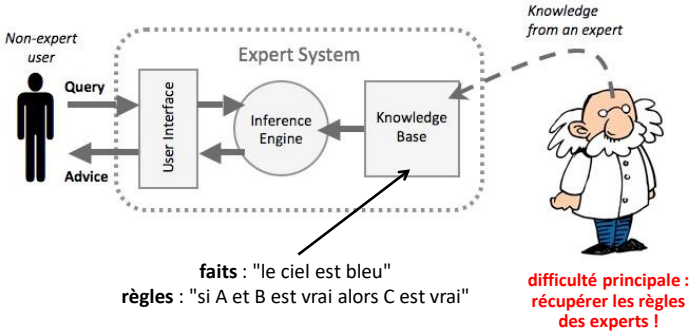
DENDRAL – composition de matériaux, Feigenbaum, Buchanan et al, 1965
PROSPECTOR – prospection géologique, 1977
R1 – Digital Equipment, 1982
DIPMETER – prospection géologique, 1982
MYCIN – diagnostique de maladies infectieuses, 1983
SOAR - Laird, 1983
Cyc – SI généraliste, Lenat and Guha, 1984
Neuron Data – Nxpert, 1985
ILOG Rules - devenu IBM Operational Decision 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 entraine Y*".

[23] 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 disparaitre.

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.





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 n'ont pas encore.

Les 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.

Il existe d'autres types de systèmes experts qui mettent en œuvre la notion de programmation par contrainte, 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.

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. 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 ces moteurs de règles (*ci-dessous* à droite). La base est notamment alimentée par l'analyse de documents disponibles sur Internet. Mais ce projet est considéré comme un échec.

Le système expert OpenCyc 4.0 qui exploitait la base de Cycorp n'est plus disponible en open source depuis 2017. Il est depuis commercialisé sous forme de licences dédiées à la recherche ou de licences commerciales.

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

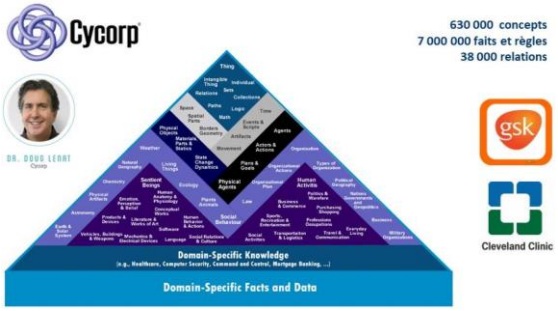
Corticon : moteur de règles sous Java et .NET, filiale de Progress Software.

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.



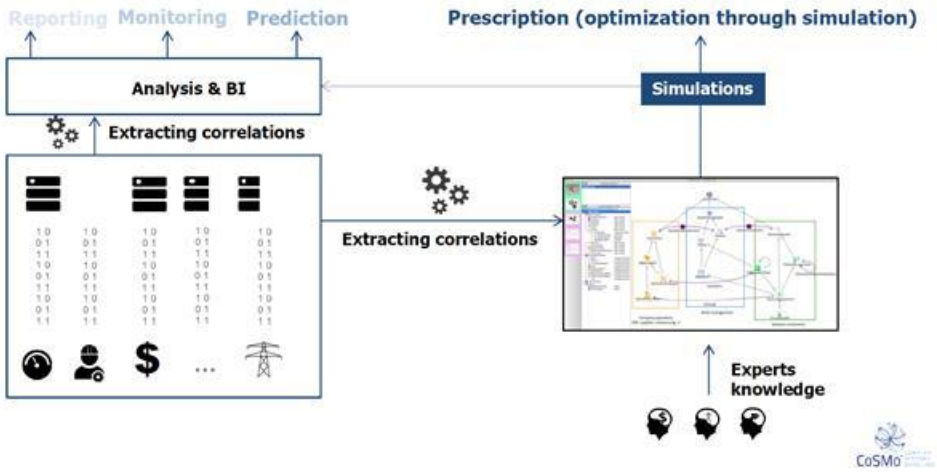
L'initiative open source **Schema.org** propose de son côté des millions de descriptions de faits exploitables par les moteurs de recherche et les moteurs de règles.

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.

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 systèmes experts peuvent aussi être alimentés par l'exploitation de données opérationnelles (big data). C'est l'approche de **Cosmo Tech** (2010, $8,2M), 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*).

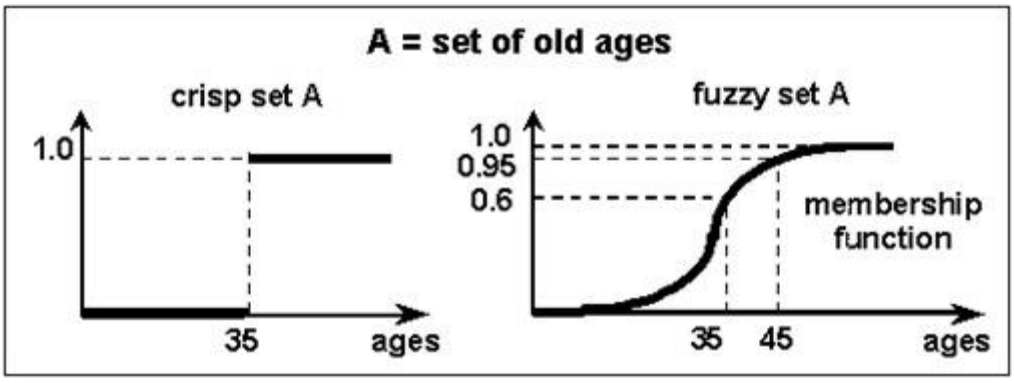
La solution est déclinée dans diverses industries comme avec leur application Asset Investment Optimization (AIO) dédiée aux énergéticiens, Crisis Management qui permet la gestion de crise et Smart Territories qui permet de modéliser des systèmes complexes pour la ville intelligente.



C'est un excellent exemple d'hybridation technologique illustrant la manière dont les systèmes experts d'intègrent dans les solutions d'IA.

La logique floue

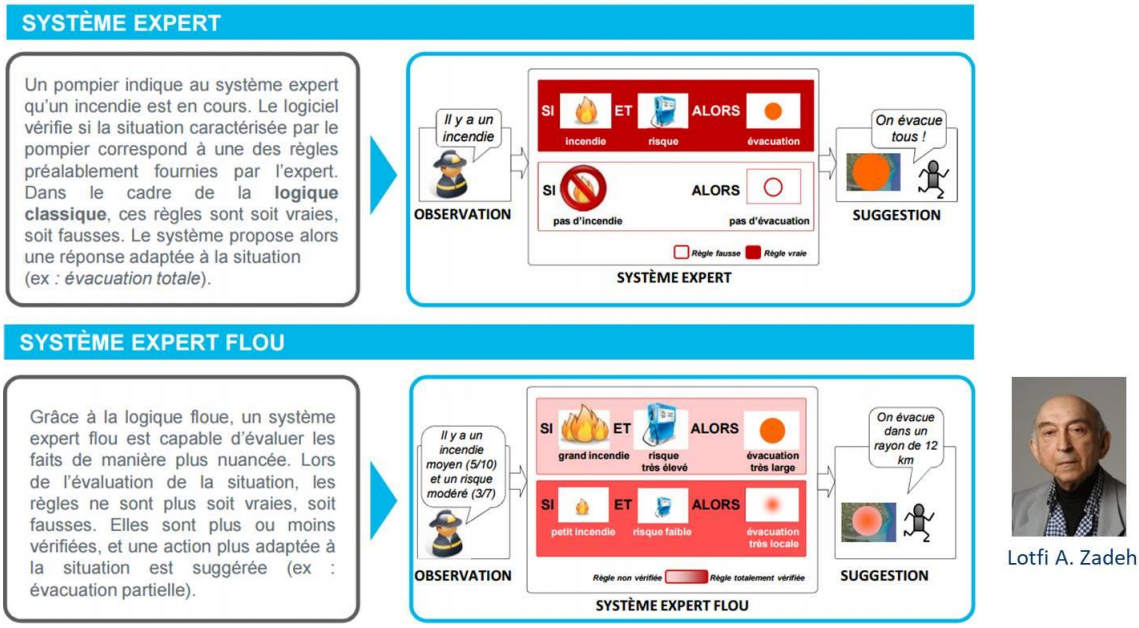
La logique floue est un concept de logique inventé par l'américain **Lofti Zadeh** ("Fuzzy Logic") en 1965[24]. Elle permet de manipuler des informations floues qui ne sont ni vraie ni fausses, en complément de la logique booléenne, mais à pouvoir faire des opérations dessus comme l'inversion, le minimum ou le maximum de deux valeurs. On peut aussi faire des OU et des ET sur des valeurs "floues".



Quid des applications ? On les trouve dans le contrôle industriel, dans des boites de vitesse chez **Volkswagen** (pour tenir compte de l'intention "floue" du conducteur), pour gérer des **feux de circulation** et maximiser le débit, 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.

[24] 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. Ca ne nous rajeunit pas !

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



Lotfi A. Zadeh

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.

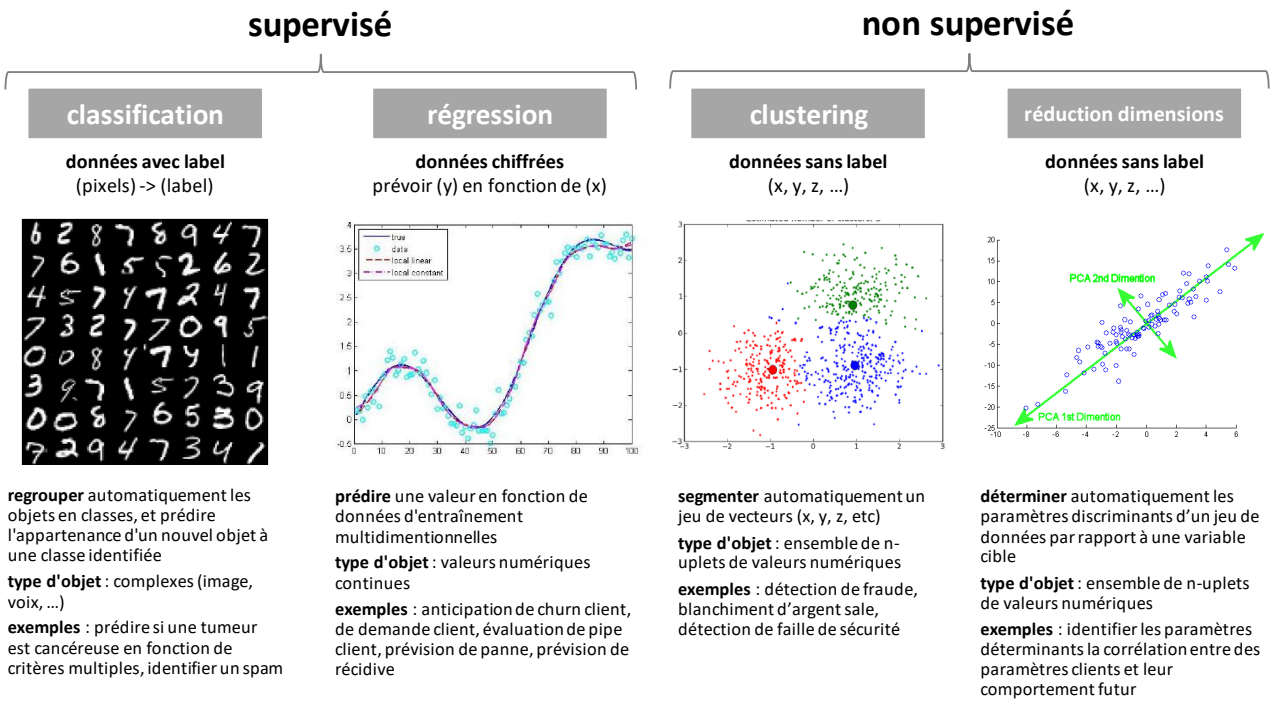
Machine learning

Le vaste domaine du machine learning, ou apprentissage automatique, vise notamment à reconnaitre des objets et à faire des prévisions à partir de données existantes. Il sert aussi à gérer des tâches de classification et de segmentation qui servent à détecter des corrélations entre paramètres 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.

En théorie, et selon son père fondateur **Arthur Samuel,** le machine learning donne aux machines la capacité d'apprendre sans être explicitement programmées. Dans la pratique, 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.

Qui plus est, l'apprentissage automatique s'appuie sur des données existantes. Elles lui permettent de produire des prévisions à partir de la généralisation d'observations. La qualité et la distribution des données d'entrainement doit permettre de faire des prévisions de bon niveau.

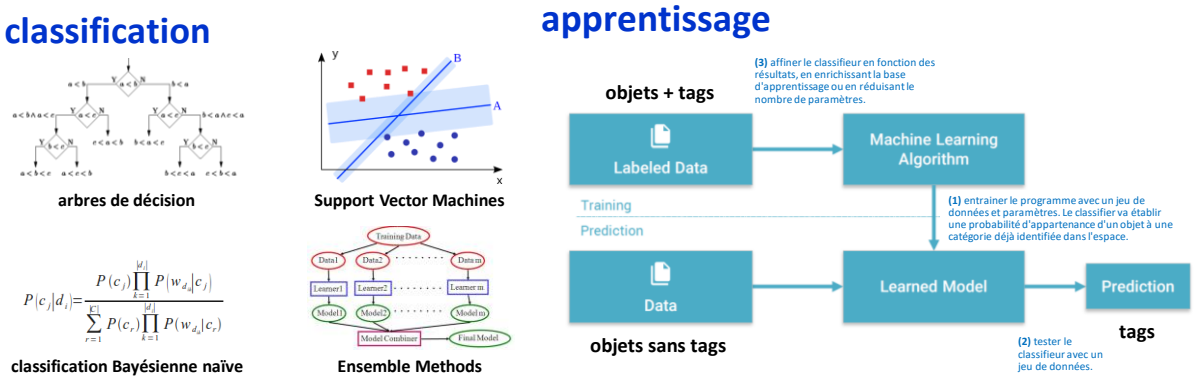
Un bon système de machine learning doit pouvoir s'adapter à différentes contraintes : une évolution permanente des données d'entrainement, ainsi que leur incomplétude et leur imperfection.



On distingue en général quatre grandes catégories de mécanismes de machine learning : la classification et la régression pour l'apprentissage supervisé et le clustering et la réduction de dimensions pour l'apprentissage non supervisé (*ci-dessus*).

Classification

Il s'agit de pouvoir associer une donnée complexe comme une image ou un profil d'utilisateur à une classe d'objet. La classification utilise un jeu de données associées à des descriptifs (les classes) pour la détermination d'un modèle. Cela génère un modèle qui permet de prédire la classe d'une donnée fournie en entrée. Exemples classiques : la reconnaissance d'un simple chiffre, l'appartenance d'un client à un segment de clients où 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 d'un logiciel.



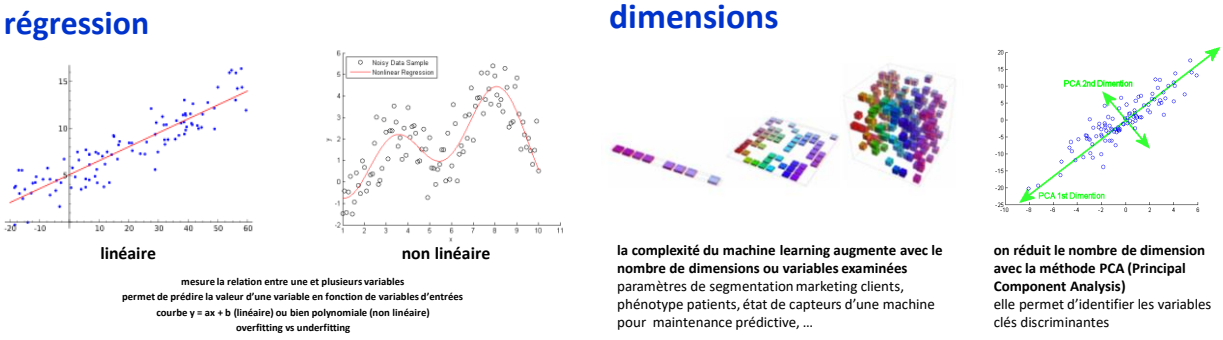
Il existe plusieurs méthodes de classification : les arbres de décision qui suivent une logique formelle, comme dans un moteur de règles, la régression linéaire ou non linéaire, les support vector machines qui cherchent à identifier une droite qui

permettra de distinguer les classes d'objets les unes des autres ainsi que des méthodes statistiques bayésiennes diverses. Un système de machine learning est entrainé avec un jeu de tests (*schéma de droite*).

Le modèle entrainé est ensuite alimenté avec de nouveaux objets pour prédire leur appartenance à une classe déjà identifiée. Dans la pratique, les spécialistes du machine learning testent différentes méthodes de classification pour identifier celle qui est la plus efficace. La plus efficace génère un maximum de bonnes réponses pour un test réalisé avec un jeu de données en entrées qui sont déjà classées mais n'ont pas servi à l'entrainement du modèle.

Régression

La régression permet de prédire une valeur y en fonction d'une valeur x à partir d'un jeu d'entrainement constitué de paires (x, y). On peut ainsi prédire la valeur d'une maison ou d'une société en fonction de divers critères. Les schémas qui illustrent ce concept utilisent uniquement une donnée en entrée et une en sortie.



Dans la pratique, 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.

Réduction de la dimensionalité

Il s'agit de déterminer dans un apprentissage non supervisé quels sont les paramètres discriminants qui ont une corrélation avec une donnée cible dans une base d'entrainement.

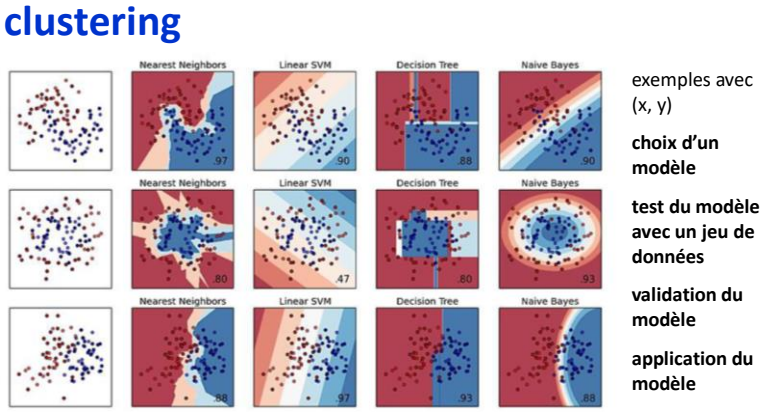
La principale méthode utilisée est la PCA, ou **Principal Components Analysis**. Cette méthode mathématique permet de réduire le nombre de variables utilisées pour prédire une valeur y en fonction d'un vecteur multidimentionnel en entrée. Par exemple, en identifiant les paramètres d'une segmentation client qui sont les plus pertinents pour prédire un comportement donné (churn, achat, …).

La réduction du nombre de valeurs utilisée va réduire la consommation de ressources machines. Mais attention, les facteurs de corrélation ne sont pas forcément des facteurs de causalité. Ces derniers peuvent être externes aux variables analysées !

La PCA est très largement utilisée dans le machine learning et le deep learning.

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 valeurs proches les unes des autres. Ce sont des clusters de valeurs.

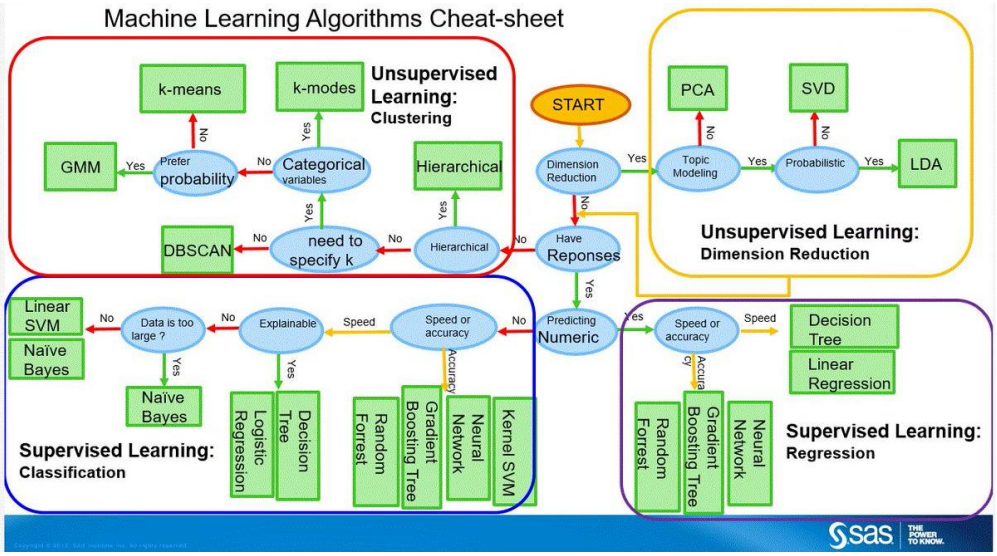


Les méthodes de clustering permettent d'identifier les paramètres discriminants de ces différents segments. Elles servent ensuite à prévoir l'appartenance à un segment d'une nouvelle valeur entrée dans le système. Là encore, si le clustering peut être automatisé, en mode non supervisé, le choix du modèle 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 clusteriser des données avec une répartition quasi-arbitraire alors que les méthodes élémentaires ci-dessus sont limitées de ce point de vue-là,

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 l'éditeur de logiciels SAS est un exemple d'arbre de décision permetant de déterminer la méthode à utiliser en fonction du problème.

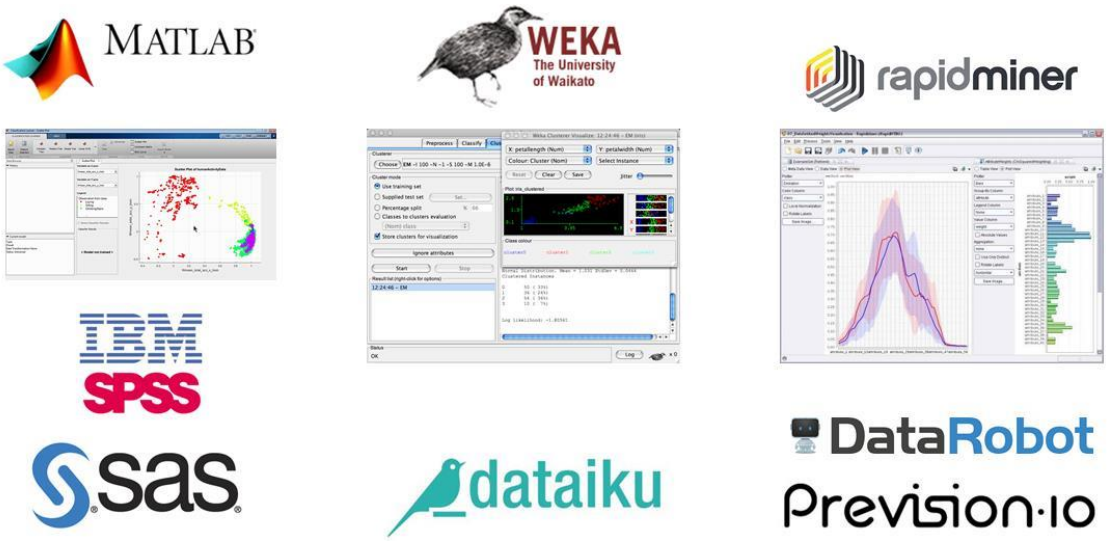


Dans la pratique, il existe un très grand nombre d'outils de machine learning. Ils combinent plusieurs types de logiciels :

- Des **langages de programmation** comme Python, Java, C++ ou autres qui sont utilisés conjointement avec des bibliothèques de calcul spécialisées dans le machine learning.
- Des **bibliothèques associées**, comme Scikit-learn, d'origine française, qui permettent de tester les modèles d'apprentissage ou d'autoapprentissage et de les mettre ensuite en production.

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

- 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**, du Suisse **Knime** (2008, $20M), de **RapidMiner** (2007, $36M)**,** les solutions de **Cognitive Scale** (2013, $40M[25]) et **CrowdFlower** (2007, $58M), le Data Science Workbench de **Cloudera** (2008, $1B) et le Data Studio du Français **Dataiku** (2013, $43M).



[25] 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.

- Des **outils d'automatisation** de la recherche de méthodes d'apprentissage comme **DataRobot** (2012, $125M, vidéo) ou le français **Prevision.io** (2016, vidéo et démonstration). Ces outils récupèrent les données du client et testent diverses méthodes d'apprentissage relevant du machine learning pour trouver celles qui sont les plus pertinentes par rapport à un objectif à atteindre. Ils parallélisent les tests de modèles en parallèle pour prédire les valeurs d'une variable dans un tableau à partir d'un tableau d'entrainement[26].

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 outils mathématiques et 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.

Dans la pratique, 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 quelques données en entrées et de générer un résultat en sortie.

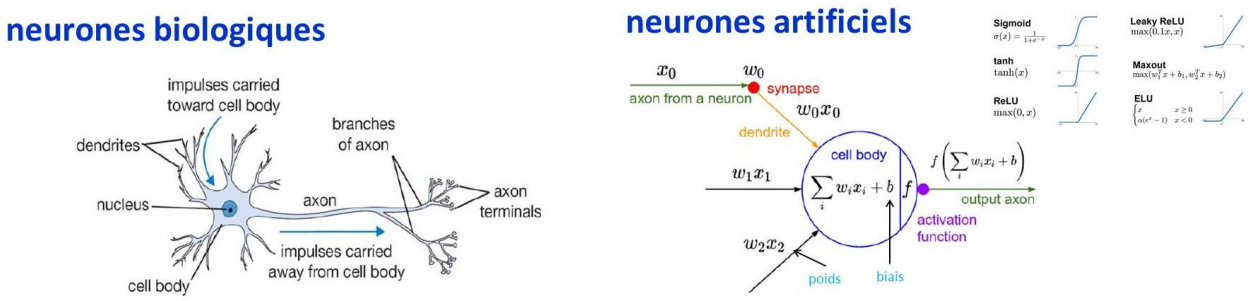
C'est une technique utilisée dans le machine learning et dans sa variante avancée du deep learning.

Les neurones artificiels

Le principe d'un neurone artificiel est de récupérer différentes valeurs numériques en entrée (x_n) associées à un poids (w_n). Le neurone artificiel moderne fait la somme des entrées multipliées par leur poids, additionne un biais (b) et lui applique ensuite une fonction qui est en général une fonction non linéaire comme une sigmoide qui génère une valeur comprise entre 0 et 1, générant une valeur % statistique facile à exploiter dans le reste du réseau de neurones. Le procédé imite vaguement le fonctionnement d'un neurone biologique qui est dans la pratique bien plus complexe[27].

[26] Les deux 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. Cf un descriptif précis du mode opératoire de prevision.io : Building a production-ready machine learning system with Prevision.io de Gerome Pistre, octobre 2017.

[27] Un neurone du cortex est généralement relié 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 gênes différents dans les neurones et par des cellules gliales qui alimentent les neurones en énergie et qui régulent la la production de neurotransmetteurs et la conductivité des axones via la myéline qui les entoure. Bref, c'est très compliqué !



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 assemble généralement 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.

Une couche cachée permet de gérer 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[28].

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 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.

Les perceptrons

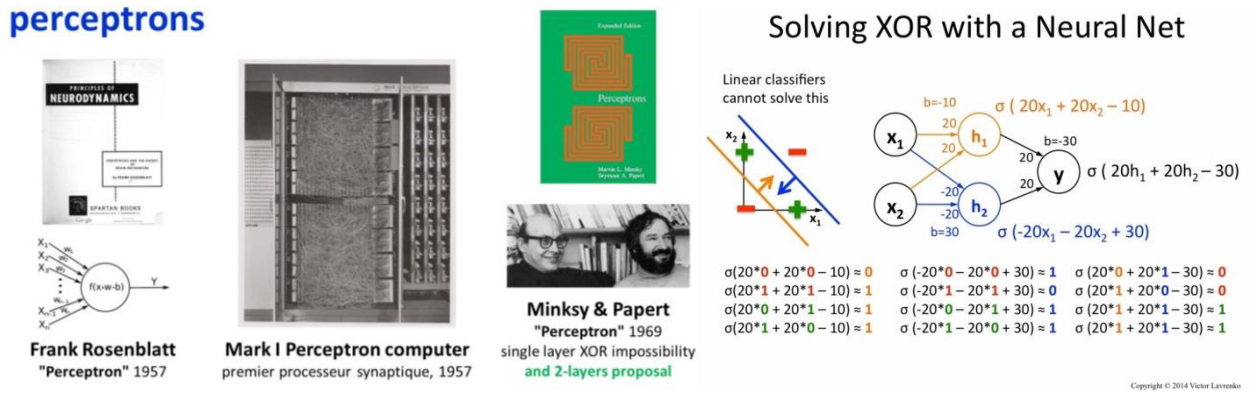
Le concept des réseaux de neurones a vu le jour en 1943 dans les travaux de **Warren McCullochs** et **Walter Pitts**. En 1949, **Donald Hebb** ajouta le principe de modulation des connexions entre neurones, permettant aux neurones de mémoriser de l'expérience.

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 !

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 zéro ou un un en sortie.

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

Le premier perceptron était 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 deux bits sont identiques ou non. 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 !

Il contribua cependant à mettre un coup d'arrêt à ces développements, le coup de grâce arrivant avec le rapport Lighthill publié au Royaume Uni en 1972. 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 reconnu toutefois son erreur d'appréciation dans les années 1980, après le décès de Frank Rosenblatt.

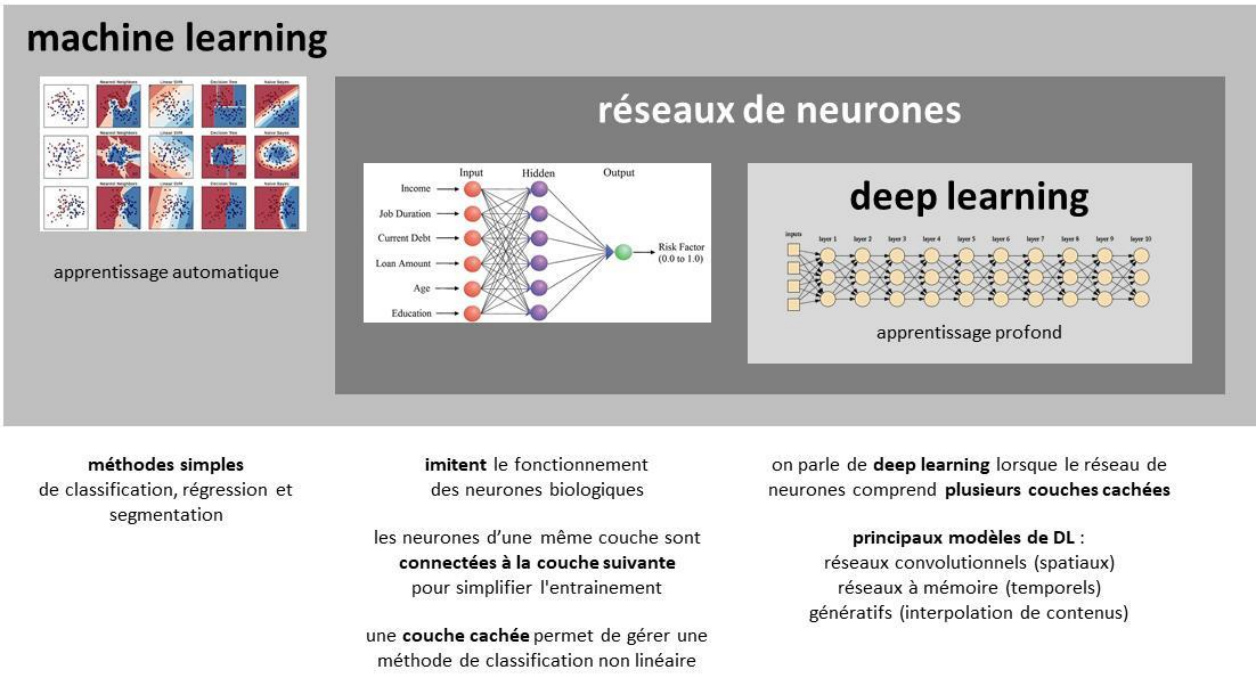
Du machine learning au deep learning

Les réseaux neuronaux ont connu ensuite un fort développement à partir des années 2000 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.

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

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

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.

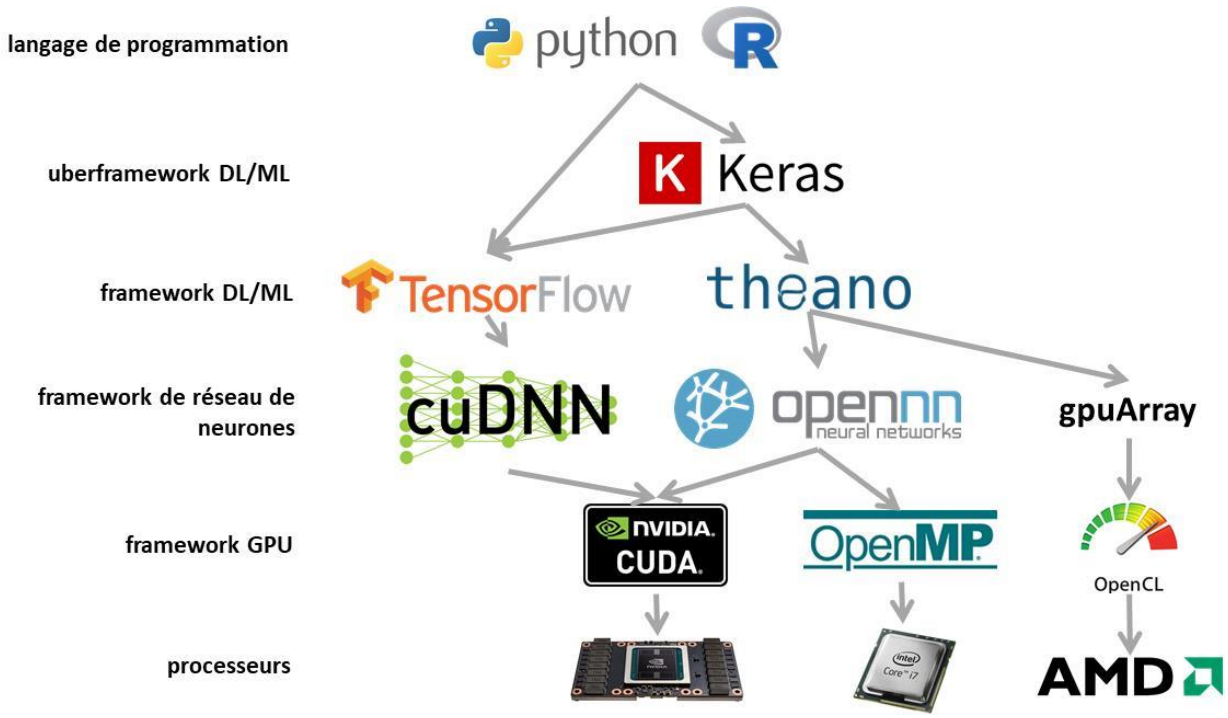


L'imagerie 2D complétée par des informations de profondeur améliorera la capacité de détection d'objets complexes. La vidéo d'un visage permettra d'améliorer la captation de la parole par l'équivalent numérique de la lecture sur les lèvres.

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.

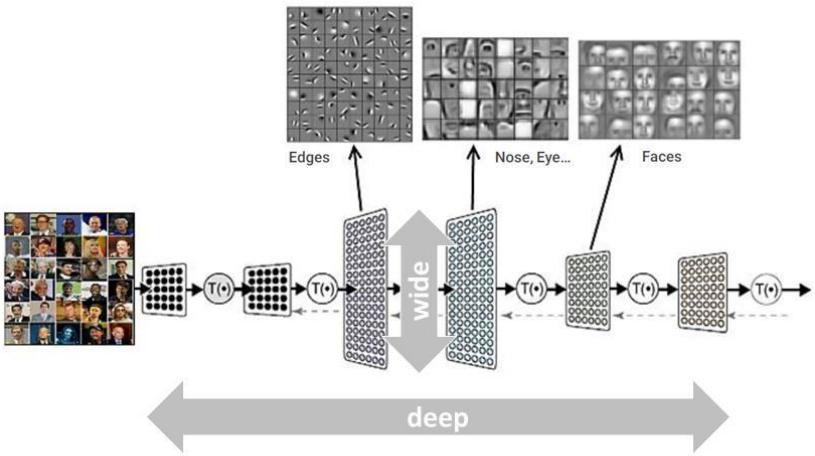
Ces bibliothèques de réseaux de neurones sont souvent exploitées elles-mêmes par des bibliothèques de machine learning ou de deep learning, comme **TensorFlow**, qui masquent la complexité du pilotage de réseaux de neurones à bas niveau et permettent par exemple de définir les modèles de réseaux de neurones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



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 ou OpenCL pour AMD, et enfin, un GPU ou un CPU au niveau matériel.

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.

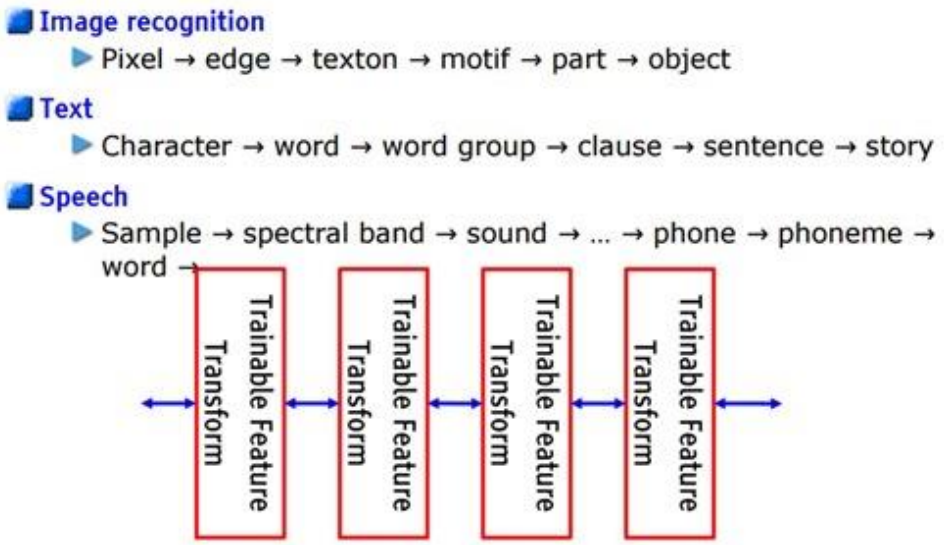


chaque niveau d'un réseau de neurones profond réalise une fonction de classification d'un type d'objet en un type d'objet supérieur (pixels => forme => nez => visage => individu)

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

Celles-ci permettent 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.

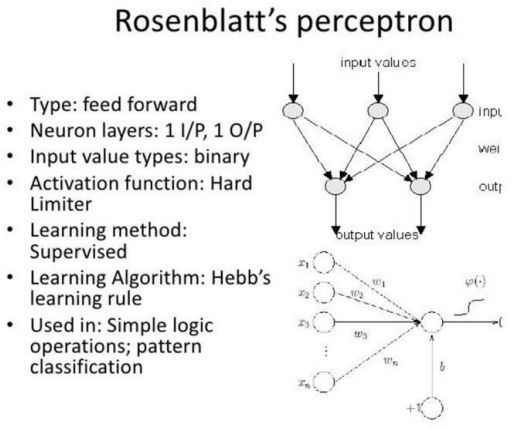
Le deep learning sert le plus souvent à reconnaître le langage, l'écriture et les images mais il peut aussi avoir d'autres usages dans les outils d'aide à la décision, dans les jeux tels que le Go avec AlphaGo et même dans l'exploitation de données structurées.



Le deep learning a d'autres usages, notamment pour 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.

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
neocognitrons 1983
error propagation 1986
recurrent neural networks 1982
restricted boltzmann machine 1986
forward propagation 199X
convolutional neural networks 1998
deep belief networks 2006
stacked autoencoders 2007
google imagenet 2012

L'histoire du deep learning a cependant véritablement démarré près de 20 ans plus tard, dans les années 1970. Il a cependant fallu attendre les années 2000 pour que l'on puisse les mettre en œuvre en pratique.

C'est sans doute dû aux progrès matériels, à la loi de Moore mais aussi aux progrès conceptuels, notamment aux travaux de Yann LeCun en 1988 et 1998 et à Geoff Hinton, particulièrement en 2006.

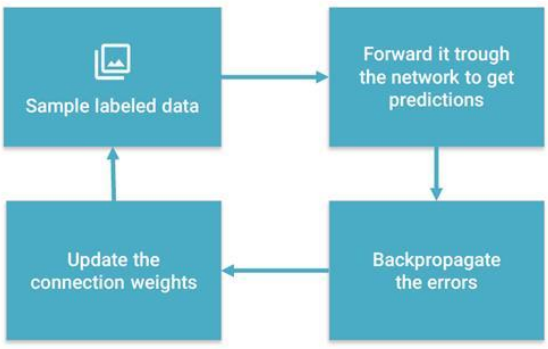
Il est de bon ton de déclarer que les chercheurs n'ont pas produit grand chose depuis et que le deep learning doit surtout aux progrès du matériel et l'abondance de données pour entrainer les systèmes. Quand on y regarde de plus près, on se rend compte qu'au contraire, les chercheurs n'ont pas cessé de faire avancer le domaine. 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.

C'est ce que nous allons vois dans ce qui suit. Ces avancées du deep learning sont étalées sur plusieurs décennies et sont continues !

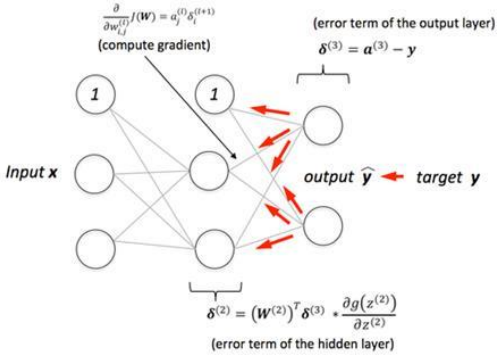
Rétropropagation d'erreurs (1969)

Elle permet l'entrainement d'un réseau de neurones couche par couche en partant du résultat et en ajustant le poids des neurones pour permettre au réseau d'identifier les objets 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.

L'apprentissage des réseaux de neurones est généralement supervisé et automatique ! Supervisé car il utilise des tags descriptifs des objets d'une base de référence et automatique car les poids synaptiques des neurones sont ajustés automatiquement.



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

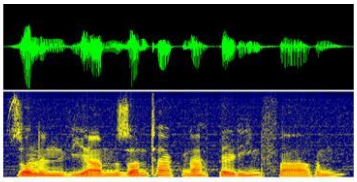
Les évolutions des méthodes de rétropropagation créées par la suite visaient surtout à économiser du temps machine car l'opération est très fastidieuse puisqu'elle doit être répétée pour chaque neurone du réseau et pour chaque objet de la base de référence. Cela donne une combinatoire très élevée !

La méthode est perfectionnée en 1986 par David Rumelhart, Geoff Hinton et Ronald Williams dans Learning representations by back-propagating errors. La plus couramment utilisée aujourd'hui est la descente stochastique de gradient, vue un peu plus loin, et qui permet d'améliorer la vitesse de convergence des réseaux lors de leur entrainement.

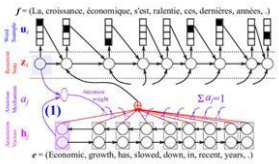
Réseaux de neurones récurrents et à mémoire (1982 puis 1993)

Ces RNN (Recurrent Neural Networks) permettent d'analyser des informations séquentielles 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.

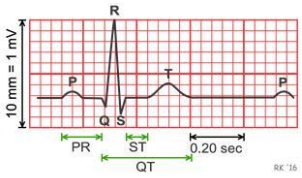
Ils peuvent aussi analyser des signaux comme le bruit de machines pour y détecter des anomalies, dans le cadre de maintenance préventive, aux prévisions de cours d'action, à l'analyse d'électro-cardiogrammes[29] et même à la détection des exoplanètes par la méthode des transits[30].



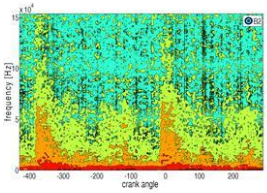
reconnaissance de phonèmes



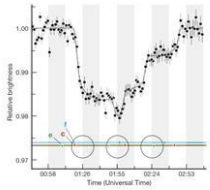
traduction



analyse d'ECG



maintenance préventive de moteur



détection d'exoplanètes

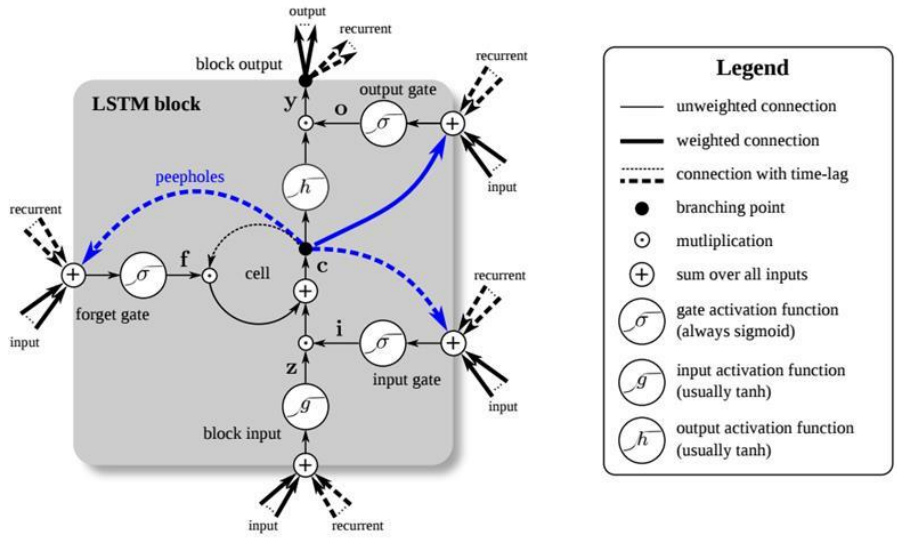
Les réseaux de neurones à mémoire ont de nombreuses déclinaisons avec notamment :

- **BPTT** (1987), BackProp Through Time, une méthode d'entrainement de réseaux de neurones récurrents.) et
- **RTRL** (1989), Real Time Recurrent Learning, une variante de réseaux de neurones récurrents.

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

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

- **LSTM** (1997), Long Short Term Memory, qui savent gérer le contexte dans lequel les contenus apparaissent[31] et sont très utilisés pour le traitement du langage et la traduction automatique.



- **GRU** (2014), Gated Recurrent Units[32], des variantes plus simples des LSTM.
- **BLSTM** (2015), Bidirectionnal Long Short Term Memory, des LSTM bidirectionnels.
- **Stacked RNN** (2015), qui sont des RNN empilés.
- **MANN** (2015), Memory-Augmented Neural Networks.

Je vous en épargne les détails, ce d'autant plus que je n'ai pas encore très bien compris leur fonctionnement dans le détail et que ces réseaux de neurones sont assez difficiles à vulgariser. Ils sont souvent combinés entre eux.

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.

Tous ces réseaux 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[33].

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 perfomantes pour la reconnaissance de la parole, la traduction automatique et les agents conversationnels réellement intelligents.

---

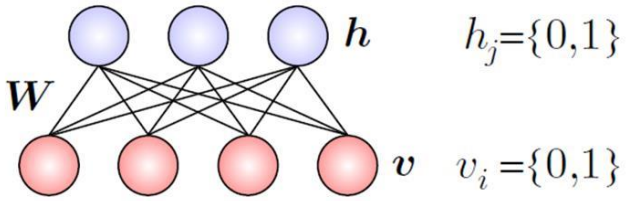
[31] Les LSTM ont été conceptualisés par Sepp Hochreiter et Jürgen Schmidhuber dans Long short-term memory, en 1997.

[32] Les GRU ont été créés par Junyoung Chung en 2014. Cf 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.

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

Machines restrictives de Boltzmann (1986)

Les DBN utilisent une seule couche de neurones source et cible et il n'y a pas de connexions entre les neurones d'une même couche.

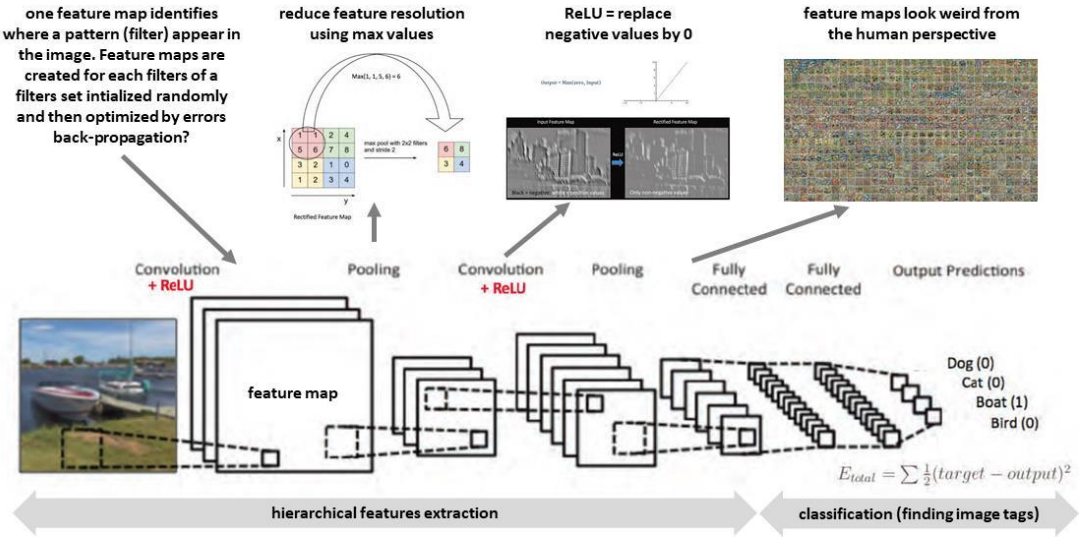


C'est le modèle le plus simple de réseau de neurones qui est ensuite exploité dans d'autres assemblages, comme les Deep Belief Networks (DBN) créés en 2006.

Réseaux de neurones convolutionnels (1989 puis 1998)

Les réseaux de neurones convolutionnels ont été inventés par le français Yann LeCun[34] et servent en premier lieu à la reconnaissance de caractères puis à la reconnaissance d'images[35]. Ce sont des outils qui servent principalement à réaliser de la classification d'objets, comme pour associer une image à une classe d'objet (chat, bateau, avion, …).

Les premiers ConvNets de production ont été déployés en 1995 pour la reconnaissance des chèques, via une solution de NCR. Les CNN, appelés aussi ConvNets (convolutional neuron networks), utilisent plusieurs techniques enchainées les unes avec les autres avec notamment des filtres et des feature maps qui consistent à identifier des formes dans les images.



[34] Yann LeCun s'était inspiré des travaux de Kunihiko Fukushima, un chercheur de la NHK, et de ses réseaux de neurones multi-couches Neocognitron. Voir Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition, 1987.

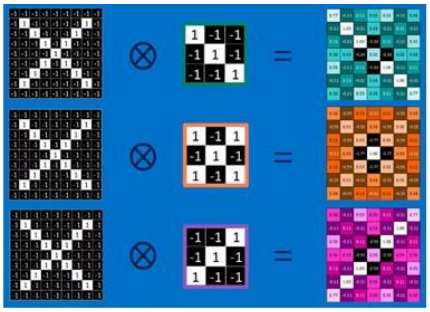
[35] Voir cette bonne explication en trois parties : A Beginner's Guide To Understanding Convolutional Neural Networks de Adit Deshpande (un étudiant aux USA), partie 1, partie 2 et partie 3, 2016.

Une feature map est une cartographie de l'apparition d'un filtre dans l'image analysée. Un ConvNet utilise un jeu de plusieurs filtres initialisé aléatoirement. Les filtres sont des matrices de quelques pixels de côté, en général 3x3 ou 4x4[36]. Les filtres 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'entrainement qui peut comprendre des millions d'images.

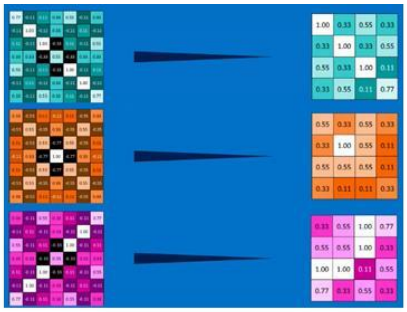
C'est très consommateur de ressources machine mais bien plus efficace qu'un simple réseau de neurones multicouches. Cela vient du fait que le réseau comprend moins de paramètres. Ce nombre de paramètres est approximativement égal à la somme de l'information des filtres de chaque convolution et des poids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

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 niveaux ou couches, chaque feature map issue d'un niveau devenant une image qui subit un traitement équivalent dans le niveau suivant.

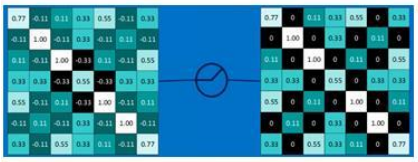
A la fin de l'histoire, la dernière couche de feature maps est reliée à une liste de tags avec une probabilité de correspondance via quelques couches de neurones dites « fully connected », à savoir que tous les neurones d'une couche sont liées à celles de la couche suivante.



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

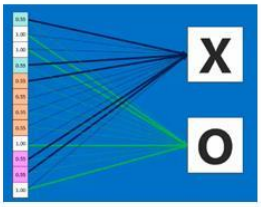


2) pooling : réduction de résolution des feature maps



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

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

C'est là qu'un chat ou un bateau sont reconnus dans l'image et que plusieurs objets peuvent être reconnus dans une même image. La dernière couche de cet empilement est un ensemble de neurones dont le nombre est égal au nombre d'objets différents à reconnaitre. Il peut être très grand mais doit rester raisonnable pour tenir compte des capacités du matériel. Ainsi, les moteurs de reconnaissance d'images n'ont-ils pour l'instant au grand maximum que quelques dizaines de milliers de classes d'objets dans cette dernière couche de réseaux de neurones[37].

Voici 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. 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 introduite 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 forcément interprétable par le cerveau humain.

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.

D'où la fameuse « non explicabilité » des algorithmes qui inquiète nombre d'observateurs[38], ce d'autant plus qu'elle se produit aussi dans les réseaux récurrents et à mémoire qui servent principalement au traitement du langage. Mais ce qui compte avant tout est la fiabilité des résultats plus que leur explicabilité. En cas de défaillance d'un réseau de neurones, l'erreur proviendra probablement d'une base d'entrainement 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'entrainement.

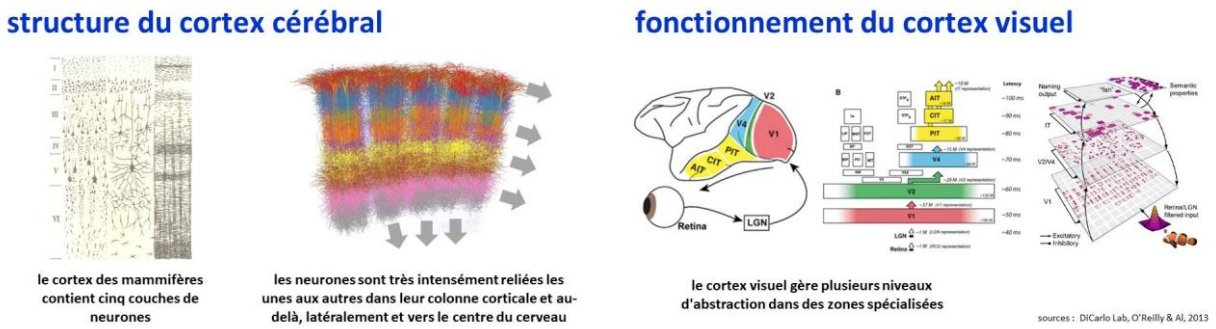
Les ConvNets s'inspirent fortement du mode de fonctionnement du cortex visuel des mammifères qui est structuré, de près, dans des colonnes corticales faites de cinq couches de neurones et qui, de loin, comprend des aires spécialisées qui élèvent progressivement le niveau d'abstraction des objets reconnus.

Par contre, contrairement au cortex humain, les ConvNets qui font de la reconnaissance d'images utilisent des représentations à très basse résolution. La majorité des ConvNets se contentent d'images comprimées à une résolution de 224x224 pixels.

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

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

L'imagerie médicale qui est plus exigeante n'y échappe pas, avec des modèles de réseaux de neurones spécialisés qui détectent d'abord les parties de l'image à analyser[39] puis les analysent avec un réseau convolutionnel à une résolution de départ de 224x224 pixels[40].



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 !

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.

Descente stochastique de gradient (2003)

Il s'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'entrainement du réseau de neurone par rétropropagation d'erreurs.

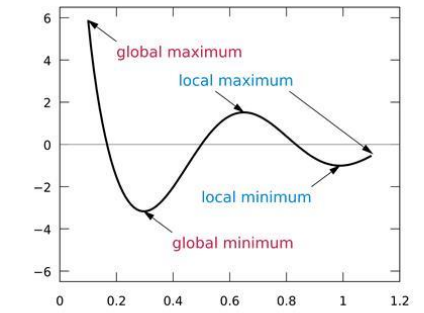
Il est important de trouver le niveau optimum global, à savoir le taux d'erreur le plus bas, et pas seulement le niveau optimum local, qui est le taux d'erreur le plus bas dans les environs du poids de départ que le réseau de neurones cherche à optimiser (cf le schéma ci-dessous qui l'explique de manière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

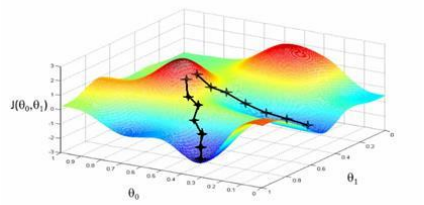
[39] Avec notamment les réseaux de neurons pyramidaux. Cf par exemple Feature Pyramid Networks for Object Detection, 2016

[40] Pourquoi 224 et pas 256 ? C'est lié à l'architecture en couche des réseaux convolutionnels et à la taille des filtres et des feature maps des convolutions, sachant qu'entre chaque convolution, la résolution des feature maps est divisée par deux. Je vous passe les détails du calcul arithmétique ! 224x224 est aussi la taille des images de la base d'entrainement ImageNet.

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'entrainement 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 de tests (en amont, des photos et en aval, des descripteurs des objets dans les photos).





dans l'apprentissage supervisé, la descente de gradient consiste à faire varier les biais des neurones pour trouver la valeur optimale globale, celle qui minimise le niveau d'erreur du réseau de neurones.

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

une des optimisations de ce procédé s'appelle "stockastic gradient descent" 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'entrainement à la fois

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

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'entrainement, ce qui génère beaucoup de calculs.

La descente stochastique de gradient est une variante de la descente de gradient qui consiste à évaluer le poids des synapses objet par objet ou groupe d'objet par groupe d'objets d'entrainement au lieu de scanner entièrement la base d'entrainement. Cela permet de réduire la quantité de calculs à effectuer et permet de converger plus rapidement vers un réseau bien entrainé.

Cette technique d'entrainement 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.

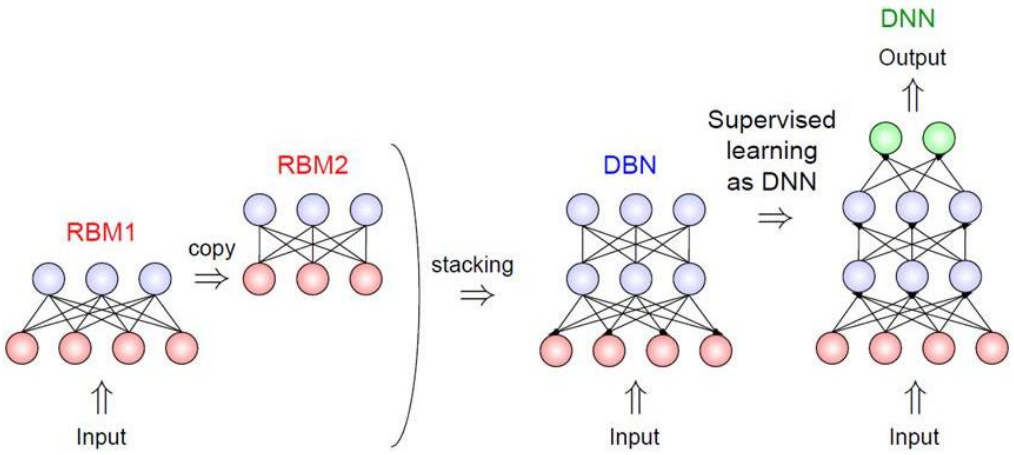
Deep beliefs networks (2006)

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 restrictives de Boltzmann empilées les unes sur les autres, étape par étape pour ce qui est de l'entrainement.

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 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) dont deux des anciens étudiants ont créé la start-up **DeepMind**, maintenant filiale de Google. Sachant que Yann LeCun était lui-même un ancien doctorant dans le laboratoire Geoff Hinton. Un bien petit monde !



Geoffrey Hinton[41] travaille pour **Google** depuis 2013, pas loin du légendaire **Jeff Dean**[42], 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[43].

Autoencodeurs empilés (2007)

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

[41] Cf 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.

[42] 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.

[43] 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 ! Cf Les Mystères mathématiques des réseaux de neurones convolutionnels, 19 février 2016.

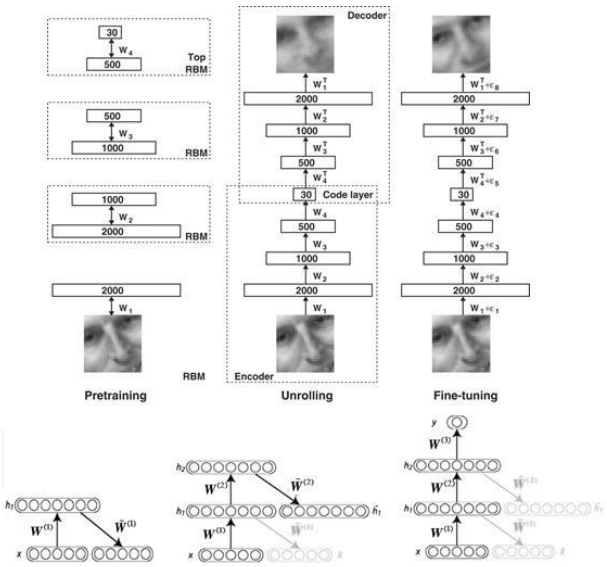
les autoencodeurs empilés sont des couches cachées de neurones qui encodent les données dans un espace dimentionnel réduit

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)

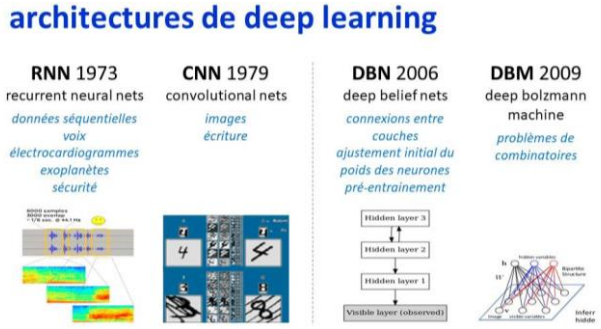


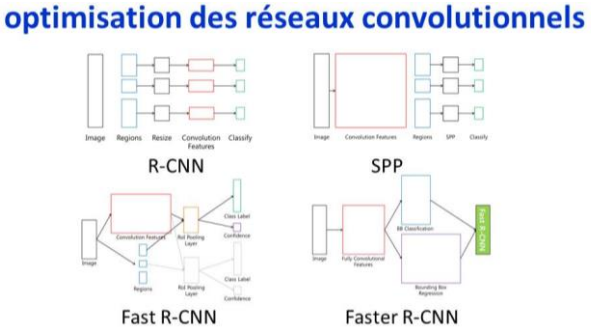
Cette technique est utilisée dans l'apprentissage non supervisé des réseaux de neurones pour identifier des variables ou fonctions cachées. Elle peut notamment servir à débruiter des images.

La méthode mathématique employée peut-être la PCA (Principal Components Analysis) que nous avons rapidement vu dans la partie sur le machine learning.

Autres méthodes de deep learning

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'entrainement 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 entrainé, il exécute ses prévisions bien plus rapidement.





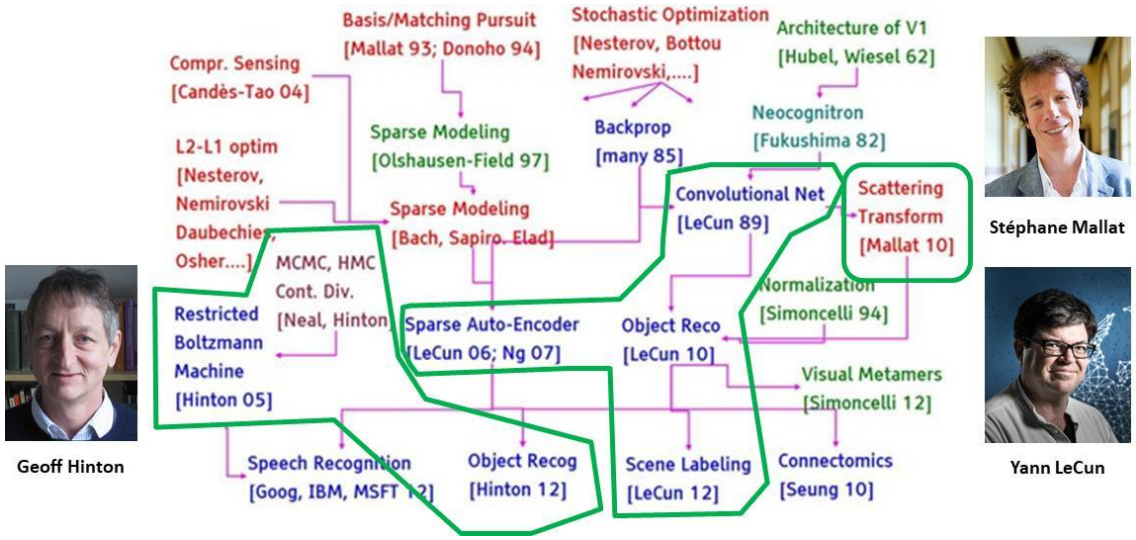
On voit aussi émerger des réseaux de **deep learning évolutifs** dont l'architecture peut évoluer de manière itérative[44].

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.

[44] Cf https://www.oreilly.com/ideas/neuroevolution-a-different-kind-of-deep-learning.

2012 est tout de même 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 entrainé 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'entrainement qui durait des semaines[45].



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'entrainement. 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.

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

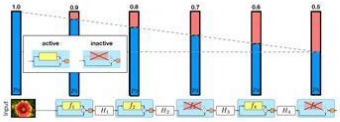
- **ResNet**[46], un réseau de neurones profonds, c'est-à-dire avec de nombreuses couches, mais qui en optimise l'entrainement, en 2015. Il sert à réduire la perte de gradients dans le processus d'entrainement. Il permettait d'atteindre un taux d'erreur de 5,7% sur la base de tests ImageNet.
- **Stockastic Residual Net**[47] (2016), qui optimise les réseaux de neurones en court-circuitant certaines couches pendant l'entrainement pour le rendre plus rapide.

[45] Cf ImageNet Classification with Deep Convolutional Neural Networks, 2012.

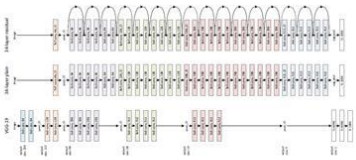
[46] Cf Deep Residual Learning for Image Recognition, de Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun, 2015.

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

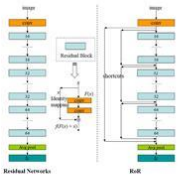
- **FractalNet**[48] (2016), qui utilise le concept des fractales pour densifier un réseau de neurones convolutionnel en répliquant certaines couches et en utilisant plusieurs circuits différents pour l'optimisation de chaque convolution.
- **DenseNet**[49] (2016), une variante des ConvNets où chaque 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 entrainement.
- **RoR**[50] (2016), une variante itérative de ResNets.
- **PathNet**[51] (2016), un réseau de neurones, chaque neurone étant un réseau convolutionnel, dont l'usage est optimisé automatiquement.
- **Mixture of Expert Layer**[52] (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.



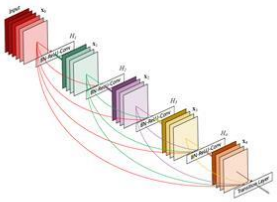
Stochastic Residual Net, 2016
accélère entrainement



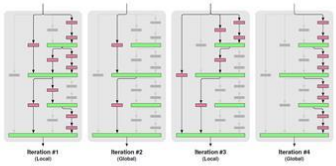
Residual Net, 2016
duplication de couches, améliore entrainement



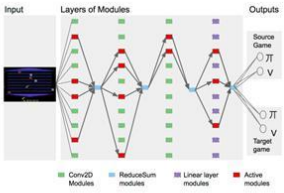
RoR, 2016
variante de ResNet



DenseNet, 2016
évite perte de gradient



FractalNel, 2016
densification de réseau



PathNet, 2016
réseau de réseau

A chaque fois, ces différents réseaux ont été entrainés avec les mêmes sources de données comme la base **ImageNet**, pour détecter 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.

---

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

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

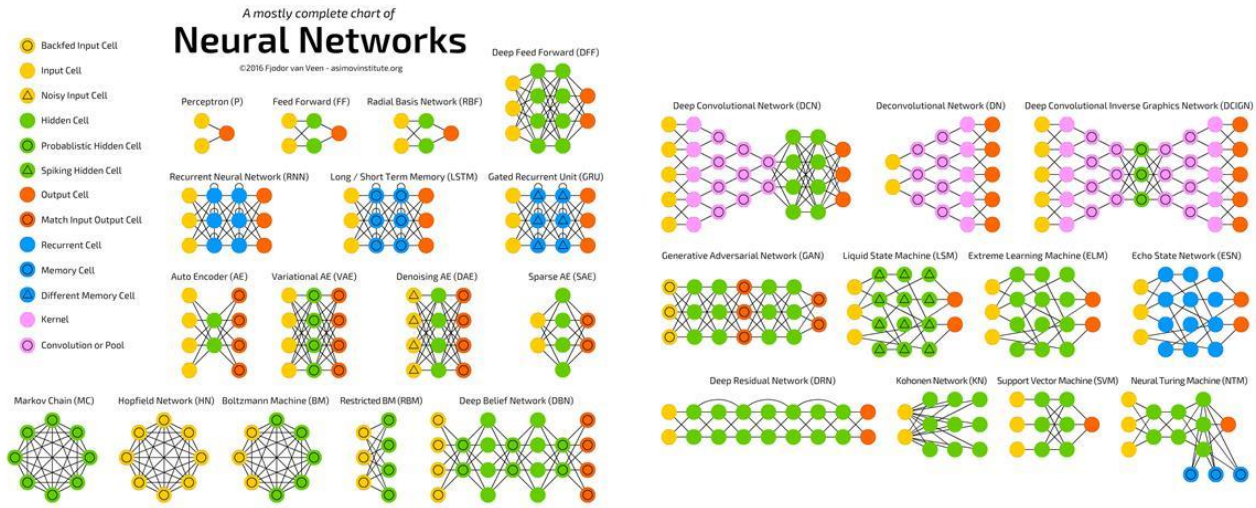
[50] Cf Residual Networks of Residual Networks: Multilevel Residual Networks, Ke Zhang, Miao Sun, Tony X. Han, Member, Xingfang Yuan, Liru Guo et Tao Liu, 2016.

[51] Cf 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.

[52] Cf 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.

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.

A ceci près que ces réseaux de neurones ont été entrainés 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[53] illustre bien leur diversité sachant que leur assemblage peut donner ensuite lieu à beaucoup de créativité en fonction des besoins.

Modèles génératifs

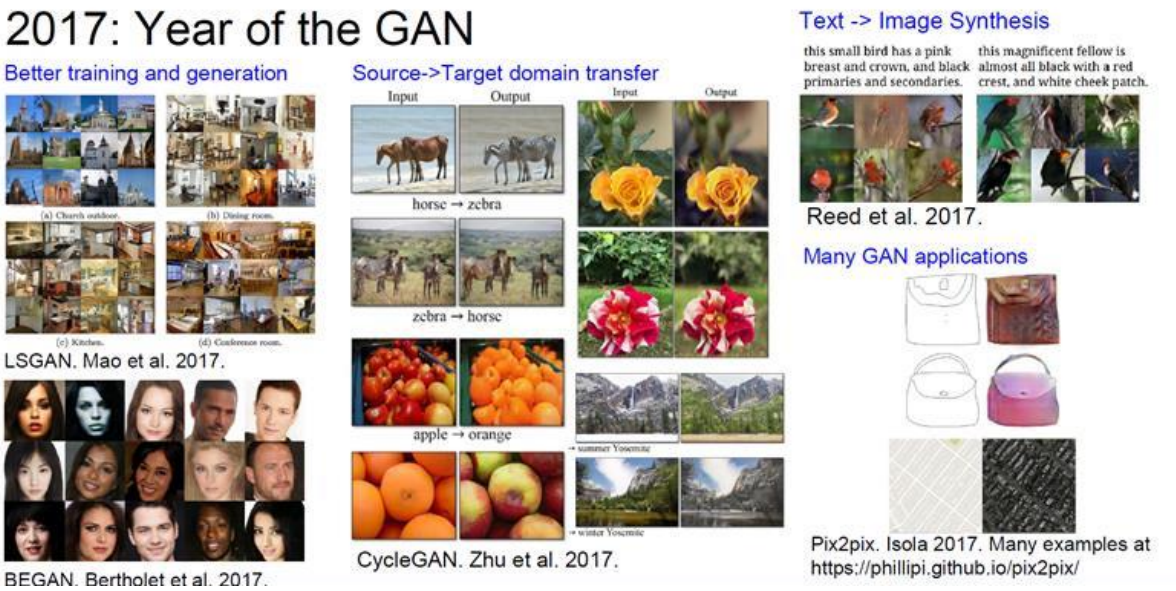
Ce sont des réseaux de neurones convolutionnels ou récurrents générant du contenu à partir de contenu existant. Ils font des prévisions, d'images vidéo suivantes d'une vidéo donnée, qui colorient des images en noir et blanc[54]. Ils peuvent aussi servir à améliorer les dialogues issus de chatbots.

Les principales techniques utilisées sont les Generative Adversarial Networks ou GANs, apparus en 2014 et perfectionnés en particulier en 2016, des réseaux de neurones non supervisés capables de générer des contenus en s'appuyant sur des générateurs à base de réseaux de convolution inversés. Ce sont des innovations toutes récentes ! Qui a dit que les algorithmes dataient tous des années 1980 ?

Je traite de cela dans la rubrique sur les modèles génératifs dans la partie vision/imagerie car c'est le principal domaine d'application des GANs.

[53] Il provient de The Neural Network Zoo de l'Institut Asimov, septembre 2016.

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



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'entrainement d'un réseau avec un jeu de données d'entrainement qui est associé à la donnée de résultat. Pour la reconnaissance d'images, il s'agit des descriptifs d'objets contenus par les images. Pour de la traduction automatique, ce sont des couples de phrases traduites d'une langue à l'autre.
- 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. Cet apprentissage ne va pas pour autant identifier automatiquement le nom des classes identifiées. On va alors utiliser un apprentissage supervisé. L'apprentissage totalement non supervisé est plus que rare.
- L'**apprentissage par renforcement** qui consiste à faire évoluer un modèle en fonction de retours externes, en général avec le monde physique. C'est une technique qui est par exemple utilisée pour optimiser le réalisme des dialogues de chatbots. Elle l'est également dans les robots qui apprennent à éviter les obstacles ou à réaliser des tâches mécaniques en tâtonnant. L'agent à entrainer 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.

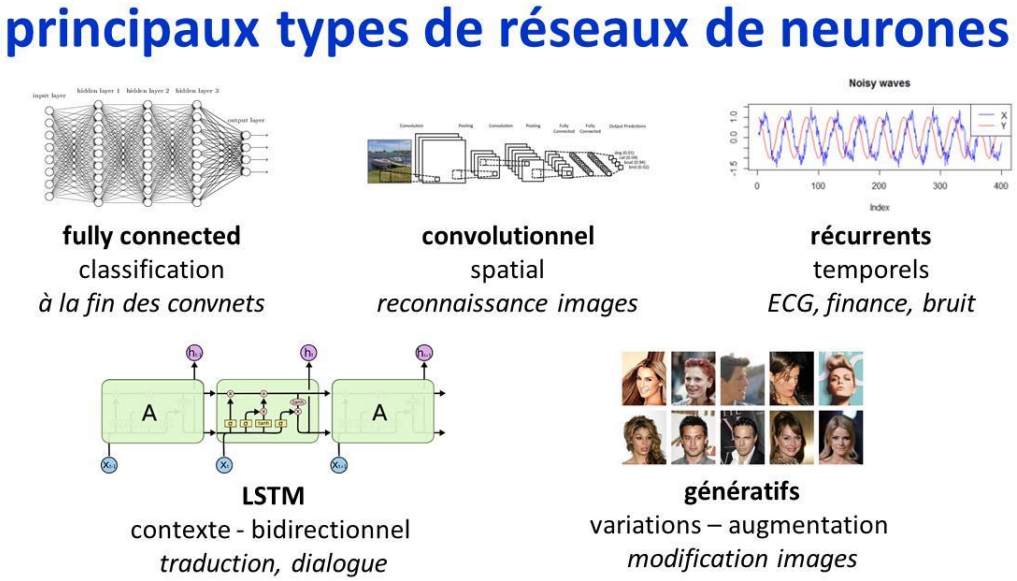
Applications du deep learning

Depuis une vingtaine d'années, 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**.

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 du temps et un bon bagage mathématique et logique ! On peut commencer par parcourir **Deep Learning in Neural Networks** de ce 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. Cela fait tout de même plus de 200 pages en corps 11 et on est largué à la cinquième page, même avec un bon background de développeur !



Il y a aussi **Deep Learning Methods and Applications** publié par Microsoft Research (197 pages) qui démarre en vulgarisant assez bien le sujet. Et puis **Artificial Intelligence A Modern Approach**, de Stuart Russell et Peter Norvig, une somme de référence sur l'IA qui fait la bagatelle de 1152 pages et qui serait l'un des B-A-BA pour les étudiants en informatique de premier cycle[55]. Mais elle commence à dater, la troisième et dernière édition étant de 2009. Il y a aussi la masse **Deep Learning** de Ian Goodfellow and Yoshua Bengio et Aaron Courville, de 802 pages[56].

[55] Le Russell et Norvig 2010 est téléchargeable gratuitement ici : https://dcs.abu.edu.ng/staff/abdulrahim-abdulrazaq/courses/cosc208/Artificial%20Intelligence%20A%20Modern%20Approach%20(3rd%20Edition).pdf . Ses auteurs prévoient de sortir une nouvelle édition en 2018 qui couvrira les avancées du deep learning.

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

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.

Du côté de la mise en œuvre pratique, le deep learning réalise des progrès continus liés au matériel et aux méthodes.

Le deep learning est très couteux en ressources marchines, surtout pendant les phases d'entrainement. 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 16000 cœurs, 170 serveurs, 20 000 catégories d'objets et 3 jours de travail[57].
- 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[58].
- En 2015, le **Lawrence Livermore Lab** créait un système gérant 15 milliards de paramètres[59] pour reconnaitre 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, entrainé sur trois serveurs en une journée[60]. 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'entrainement[61].

Mais ce sont des cas extrêmes, les réseaux de neurones en production courante ayant généralement un nombre plus raisonnable de paramètres, compris entre des milliers et des dizaines de millions.

---

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

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

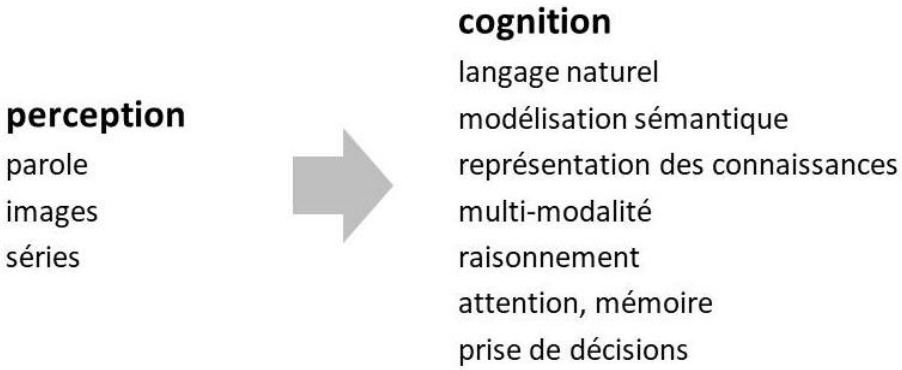
[59] Cf Large-scaled deep learning ont the YFCC100M dataset, 2015.

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

[61] Cf Deep learning image classification, Nvidia, 2016.

Toutes ces performances vertigineuses s'expliquent notamment par la vague de l'usage de GPU et de processeurs neuromorphiques dont la structure interne est plus proche 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. Il est aussi fort probable que les ordinateurs quantiques, toujours en devenir, joueront un rôle important pour augmenter la puissance de certains types de réseaux de neurones, mais pas forcément les RNN et les ConvNets.

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. Dans la pratique, ces techniques dites cognitives sont des techniques avancées de traitement du langage avec une vision plus statistique que logique.



Outils du deep learning

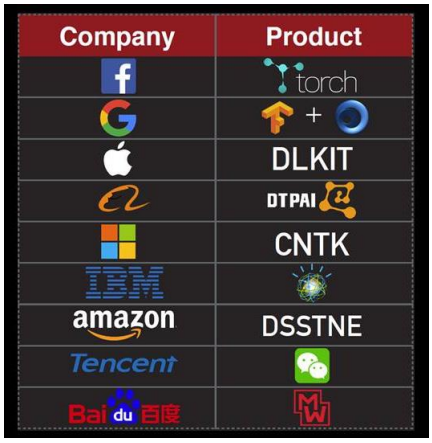
Poursuivons cette partie sur le deep learning en évoquant l'offre des outils de création des solutions les mettant en œuvre. Il s'agit d'outils de développement exploitant des langages déclaratifs comme Python. Ils permettent de créer des modèles de réseaux de neurones avec leurs différentes couches.

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 entrainement avec une boucle de programme qui va scanner un jeu de test taggé et faire de la rétropropagation de gradient dans le réseau de neurones[62].

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.

[62] Cet article décrit la prise en main avec quelques vidéos bien choisies ainsi que les cours de Stanford associés : http://blog.cremedelacreme.io/2017/06/01/digital-ressources-deep-learning/.

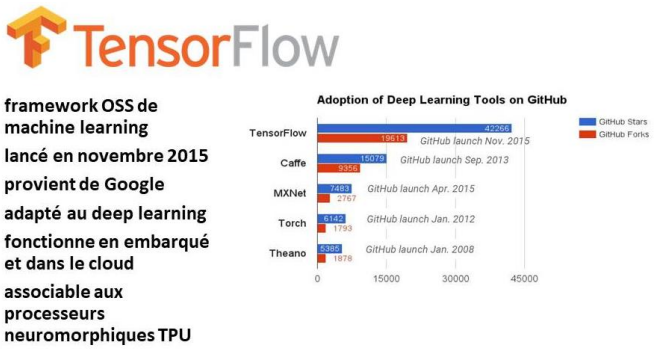
Les grands acteurs du numérique proposent tous leurs frameworks et outils open source de création de réseaux de neurones : TensorFlow chez Google, Torch chez Facebook, Cortana NTK chez Microsoft, la plateforme Watson chez IBM ou encore DSSTNE chez Amazon. Mais les startups ne sont pas en reste, comme **Theano** ou **H2O**.



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. Ca tombe bien car il sert aussi à développer la partie back-end de nombreux sites web.

L'un des frameworks sort du lot, tout du moins côté usage chez les startups : **TensorFlow** dont le développement a été initialisé par Google. Il fonctionne en embarqué aussi bien que sur serveurs et dans le cloud. 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, ce qui explique qu'il ressorte du lot dans la petit comparaison *ci-dessous à droite* (source).

Le nom TensorFlow vient de Tensor qui décrit les matrices multidimentionnelles gérées par le système. Google a annoncé au Gogle 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 voir 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.

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).

1- définition d'un modèle de CNN

```
  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')
```

convnet 1

pooling 1

normalisation 1

```
  # 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')
```

convnet 2

normalisation 2

pooling 2

Il faut ensuite puis à programmer son entrainement, ce qui peut requérir de l'optimisation programmatique, puis son exécution en mode run-time (ci-dessous, toujours en Tensorflow).

2- entrainement

```
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)
```

propagation avant et gradients

```
  # 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
```

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'entrainement.

De son côté, **Theano** est un projet académique lancé par l'Université de Montréal. Il est très bien supporté et apprécié pour sa rapidité de fonctionnement. Il est aussi assez couramment utilisé dans les startups. Mais TensorFlow a pris le dessus depuis 2016.

Certains outils sont exploités de manière combinée. Ainsi, la bibliothèque de prototypage de deep learning **Keras** peut-elle s'appuyer sur TensorFlow ou Theano.

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

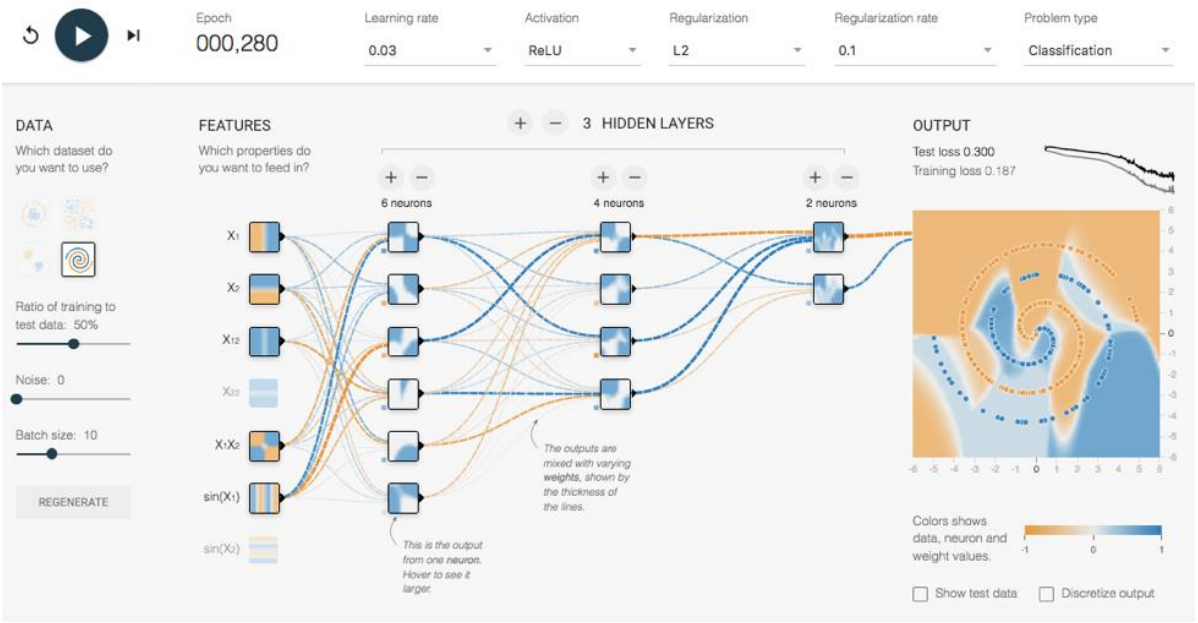
Voici quelques-unes des solutions de deep learning les plus courantes pour les développeurs de solutions extraites de l'édition 2017 du Guide des Startups.

Outil	Usage
IBM Watson	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 ssez couramment utilisé par les startups, IBM étant très actif dans leur recrutement.
scikit learn	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 (prévision) 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.
TensorFlow	TensorFlow est une bibiothèque 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.

theano	Theano est un projet issu de l'Université de Montréal. C'est une bibliothèque pour Python qui peut s'exécuter sur CPU ou GPU et sert à exploiter des matrices pour faire du deep learning. Le système peut convertir des formules mathématiques complexes en langage C pour optimiser le temps de calcul.
K	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 à base de réseaux de neurones. 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.
Microsoft CNTK	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 ls applications de machine learning et de deep learning.
amazon	Amazon propose DSSTNE, une bibliothèque en cloud pour la création de solutions de recommandation. Amazon propose aussi un grand nombre de bibliothèques pour le traitement du langage, la génération de parole synthétique et la reconnaissance d'images.
wit.ai	Wit.ai est un framework de traitement du langage originaire de Facebook, ou plus précisément, d'une acquisition de Facebook qui date de 2015. La startup avait été créée à San Francisco par un trio de français, Alexandre Lebrun, Willy Blandin et Laurent Landowski.
clarifai	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.
PyBrain	PyBrain est une bibliothèque de réseau de neurones bâtie en Python.
VW	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 exploiter des CPU multicoeur et est compilée à partir de C++.
torch	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 qui est un langage plus populaire. PyTorch est apprécié pour le prototypage de solutions de deep learning.
Caffe2	Caffe2 est un framework open source générique de deep learning. La première version provenait de Berkeley 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 entrainés sont aussi facilement déployables sur architectures mobiles.
H_2O.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.

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.

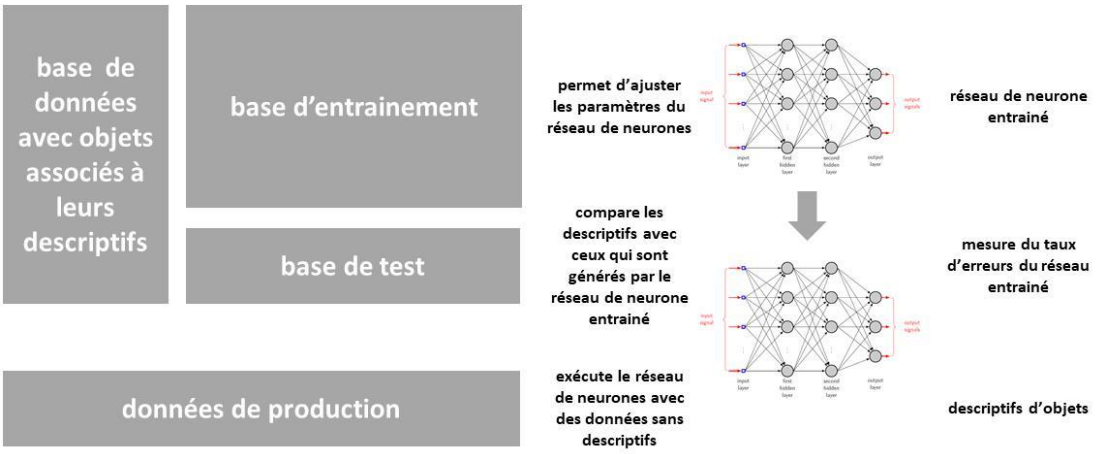
Vous pouvez aussi simuler des réseaux de neurones simples avec cet outil exploitant **TensorFlow** (exemple *ci-dessous*). TensorFlow peut en effet servir aussi bien à gérer un réseau de neurones multicouches de deep learning tout comme des solutions plus simples de machine learning, à base ou pas de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



Données

Reste à évoquer un autre point important pour le machine learning et le deep learning : l'origine et la qualité des données !

On distingue généralement trois types de données pour entrainer un système de machine learning et de deep learning : les données d'entrainement, les données de test et les données de production.



Les données d'entrainement et de tests contiennent leur descriptif, à savoir, l'information qui doit être générée par le système à entrainer. C'est un jeu de test doté d'une bonne représentation spatiale de l'espace du possible de l'application.

On le découpe arbitrairement en deux sous-ensembles, l'un pour l'entrainement et l'autre pour les tests d'erreurs du réseau de neurones entrainé. En général la part de la base taggée dédiée à l'entrainement est plus grande que celle qui est dédiée aux tests.

Les données d'entrainement et de tests sont indispensables pour la grande majorité des systèmes d'IA dits supervisés. D'ailleurs, les systèmes dits non supervisés ont aussi besoin au départ de données taggées, même si certaines étapes de l'entrainement sont non supervisées, comme pour segmenter une base.

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'entrainement qui permettent d'ajuster celui d'un réseau de neurones déjà entrainé.

Données d'entrainement

Ce sont les jeux de données qui vont servir à entrainer 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'entrainement du système, mais plus il sera long. Si vous n'avez pas de données déjà taggées pour entrainer un modèle de machine learning ou deep learning, vous n'irez pas bien loin !

Les bases d'entrainement d'images ont une taille qui dépend de la diversité des objets à détecter. Dans l'imagerie médicale, des bases d'entrainement de pathologies spécialisées peuvent se contenter de quelques centaines de milliers d'images pour détecter quelques centaines de pathologies. A l'autre extrémité de la complexité, la base d'entrainement d'images de Google Search s'appuie sur plus de cent millions d'images et permet la détection de plus de 20000 objets différents.

L'entrainement d'un système de 50 000 images dure au minimum un quart d'heures dans des ressources en cloud. Lorsque l'on passe à des centaines de millions d'images, il faudra des milliers de serveurs et jusqu'à plusieurs semaines pour l'entrainement !

Dans la pratique, les jeux d'entrainement de solutions de deep learning sont limités en taille par la puissance de calcul nécessaire. Je ne sais pas dire si cet entrainement peut fonctionner de manière incrémentale au gré de l'ajout de données où si il faut relancer un batch de calcul complet. Les techniques de machine learning traditionnelles doivent pouvoir à priori exploiter de plus gros volumes de données.

Il est évidemment nécessaire de disposer de données d'entrainement de qualité, ce qui nécessite souvent un gros travail de filtrage et de nettoyage préalable.

Données de test

Ce sont les données, également taggées, qui serviront à vérifier la qualité de l'entrainement d'un système. Ces données doivent avoir une distribution voisine des données d'entrainement, au sens où elles doivent être bien représentatives de la diversité des données que l'on trouve dans la base d'entrainement et que l'on aura dans la base de production.

Dans la pratique, les données de tests sont un sous-ensemble d'un jeu de départ dont une partie sert à l'entrainement et une autre partie, plus limitée, sert aux tests. Elles seront injectées dans le système entrainé 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.

Données de production

Il s'agit des données non taggées qui alimenteront le système lors de son utilisation en production pour faire une prévision des tags manquants.

Alors que les données d'entrainement sont normalement anonymisées pour l'entrainement du système, les données de production peuvent être nominatives ainsi que les prévisions associées générées par la solution.

La nouvelle réglementation GDPR 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 réentrainement.

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 maitriser. 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.

Les données d'entrainement des systèmes d'IA doivent être bien taggées, soit automatiquement, soit manuellement. De nombreuses bases de référence d'images taggées l'ont été via de la main d'œuvre recrutée en ligne via des services du type d'**Amazon Mechanical Turk**[63].

Là encore, les entreprises et les startups devront prendre en compte le règlement européen GDPR 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[64] !

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'entrainement des modèles.

---

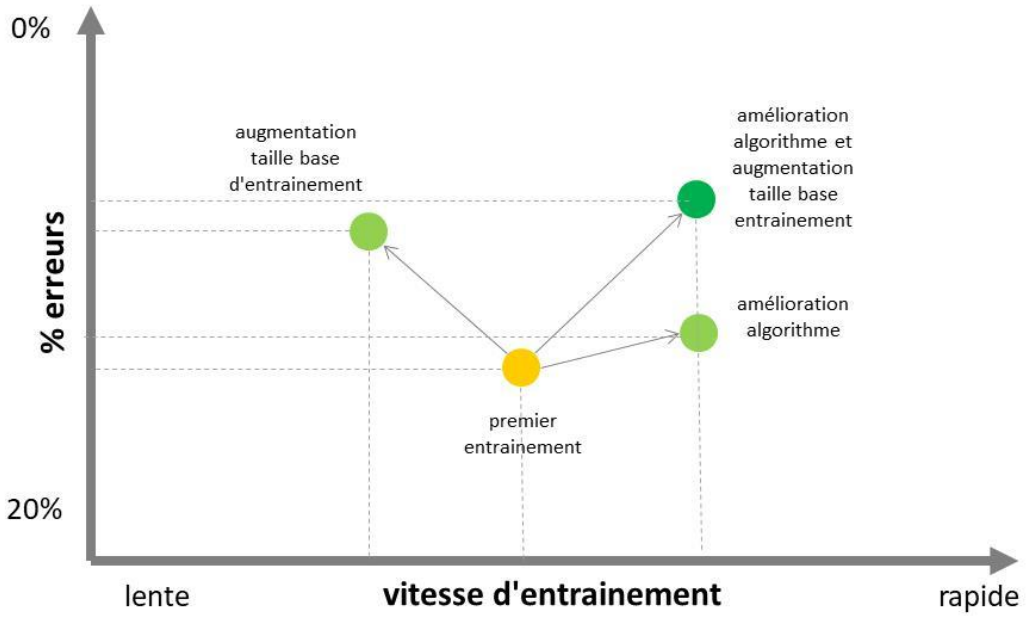
[63] Cf Inside Amazon's clickworker platform: How half a million people are being paid pennies to train AI, de Hope Reese et Nick Heath, 2016.

[64] Lorsqu'un réseau de neurones aura été entrainé 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 entrainé avec. Mais les données utilisées dans l'entrainement 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 entrainé. Influence qui est normalement négligeable à l'échelle d'un seul utilisateur. A l'envers, un réseau de neurones bien entrainé 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.

Pour ce qui est de la reconnaissance des images, il faut distinguer le temps d'entrainement 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 entrainer un modèle. Si l'algorithme utilisé n'est meilleur que dans la vitesse d'entrainement, 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 entrainé. Par contre, avec plus de données d'entrainement, celui-ci sera plus long.

Donc, comme illustré dans mon petit schéma *ci-dessous*, il faut à la fois de meilleurs jeux de données et de meilleurs algorithmes pour que l'entrainement soit aussi rapide que possible. C'est notamment utile pour réduire la consommation énergétique de l'IA. Bref, pour faire de l'IA verte[65] !



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 !

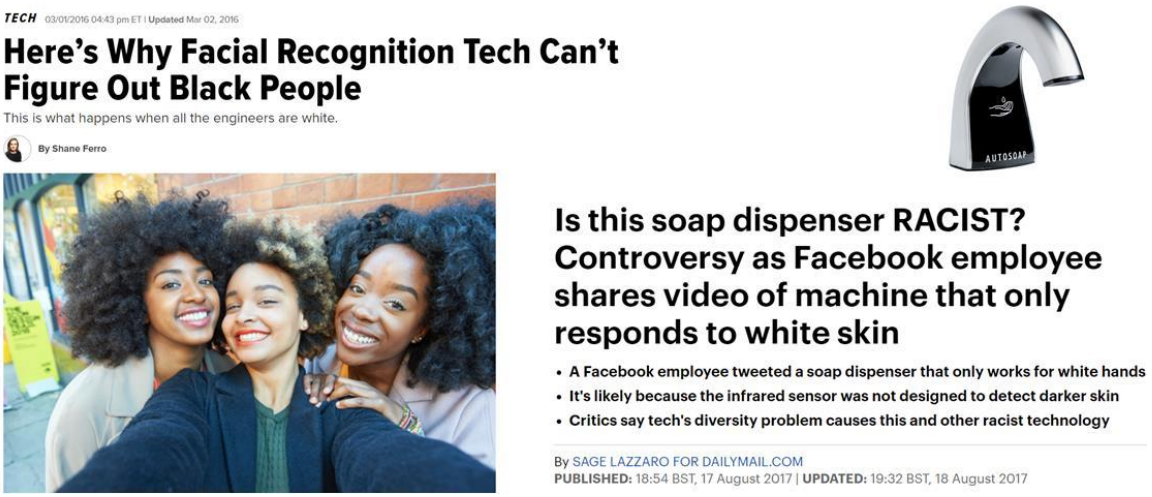
[65] C'est un des objectifs du chercheur **Stéphane Canu** de l'INSA Rouen qui planche sur l'optimisation de gros modèles de vision artificiel et de traitement du langage. D'où le projet de recherché collaborative "Deep in France" lancé par différents laboratoires et financé par l'ANR.

Biais des algorithmes et 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[66].

Deux annecdotes l'illustrent parfaitement : chez Facebook, les femmes ingénieures de couleur se sont rendu compte 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'entrainement 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.



Une IA doit donc être alimentée par des jeux de données d'entrainement 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[67].

En termes statistiques, cela veut dire que les données doivent avoir un fort écart type et une distribution similaire à celle du marché visé.Les données d'entrainement d'IA qui portent sur le fonctionnement de machines doivent répondre aux mêmes exigences.

Ainsi, si on entraine une IA à reconnaitre le bruit de moteurs en panne, il faut disposer d'une base d'entrainement 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.

---

[66] Cf l'ouvrage de référence Weapons of Math Destruction de Cathy O'Neil et cette vidéo d'une heure où elle résume son propos.

[67] Cf Forget Killer Robots—Bias Is the Real AI Danger de John Giannandrea (Google), octobre 2017,

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.

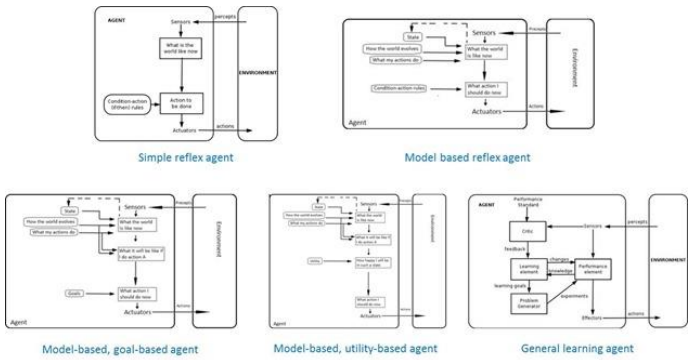
Dans la pratique, les solutions d'intelligence artificielle sont des agents ou des réseaux d'agents ! Les réseaux d'agents fonctionnent de manière coordonnée et collective. La coordination de réseaux d'agents est un domaine scientifique à part entière.

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.

Les solutions d'IA sont des réseaux d'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.

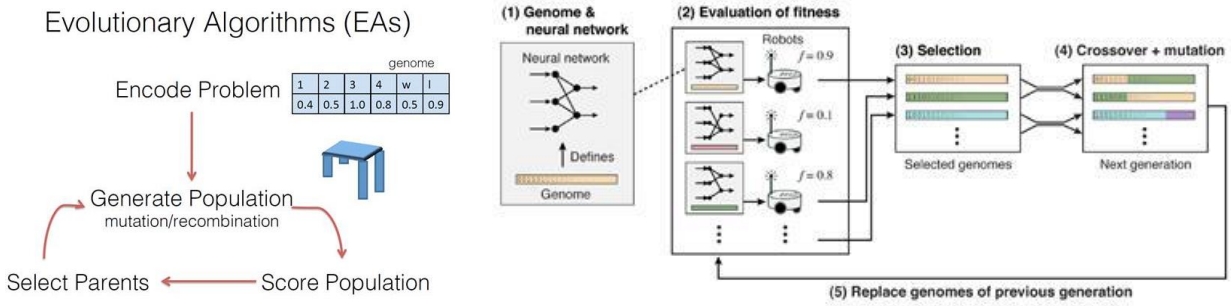
Cette intégration est un savoir nouveau à forte valeur ajoutée, au-delà de l'intégration traditionnelle de logiciels via des APIs classiques. Cette intelligence artificielle intégrative est à l'œuvre dans un grand nombre de startups du secteur et en particulier dans celles de la robotique.

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.

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.



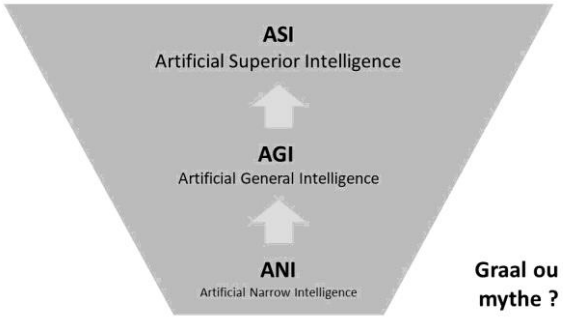
Ces systèmes exploitent des algorithmes évolutifs qui sont à la croisée des chemins du deep learning et des réseaux d'agents. Ils consistent à tester différentes combinaisons de réseaux de neurones voire de réseaux d'agents les intégrant pour comparer leur efficacité et conserver les variantes les plus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

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

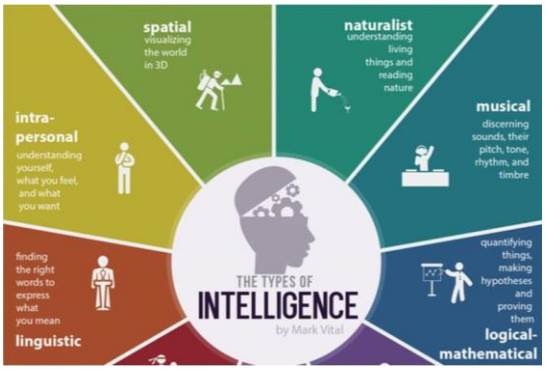
Au plus haut niveau conceptuel, on segmente l'IA en **IA forte** qui imiterait le cerveau humain avec une conscience et **IA faible**, qui évoluerait de manière incrémentale à partir d'outils plus élémentaires.

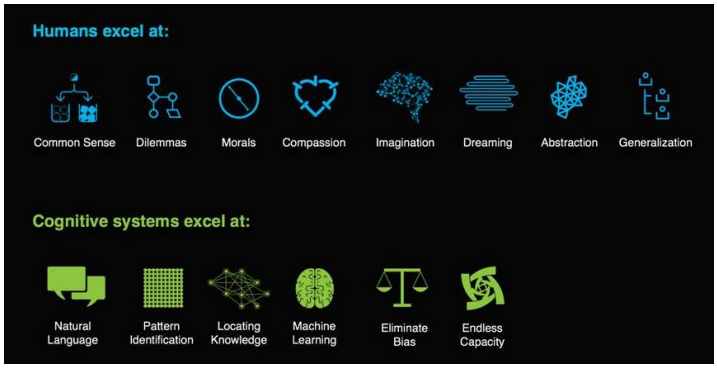
La distinction entre IA forte et IA faible se retrouve dans cette **classification** de la portée de l'IA avec trois niveaux d'IA : l'ANI, l'AGI et l'ASI.



L'**Artificial Narrow Intelligence** (ANI) correspond à la capacité de traitement de problèmes dans un domaine précis. C'est l'état de l'art actuel, exploitant aussi bien le machine learning que le deep learning ou les systèmes experts. 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, Apple SIRI, Amazon Alexa, Microsoft Cortana et Google Now.

Si l'IA n'imite pour l'instant pas encore 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. Certains estiment qu'elle ne comprendrait qu'un Go de données !





L'**Artificial General Intelligence** (AGI) correspond conceptuellement au niveau d'intelligence équivalent à celui de l'Homme, avec un côté polyvalent, avec la capacité à raisonner, 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 de concepts, la perception du monde et de soi-même, l'invention et la créativité, la capacité à réagir à l'imprévu dans un environnement complexe physique comme intellectuel ou encore la capacité d'anticipation[68]. Et à plus haut niveau, il faut intégrer la conscience, les sentiments, la sagesse et la connaissance de soi.

On pourrait y ajouter la capacité à ressentir des émotions personnelles ou sentir celle des autres (l'empathie), avoir des envies et des désirs et aussi savoir gérer ses pulsions et agir avec plus ou moins de rationalité. 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. Il reste aussi à développer des solutions d'IA capables de créer des théories et de les vérifier ensuite par l'expérimentation.

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"[69], "*L'économie a besoin d'intelligence, pas de conscience*" ! Laissons donc une partie de notre intelligence voire une intelligence

[68] 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..

[69] Intervenant en juin 2016 dans la **conférence USI** organisée par Octo Technology à Paris.

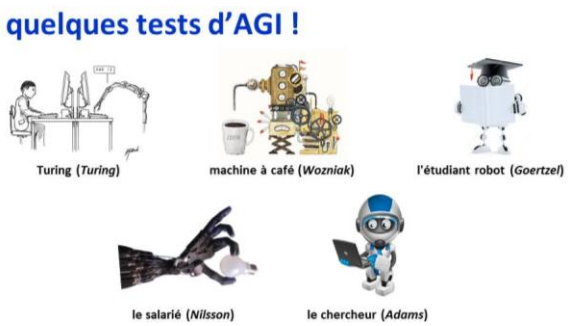
plus développée aux machines et conservons la conscience, les émotions et la créativité !

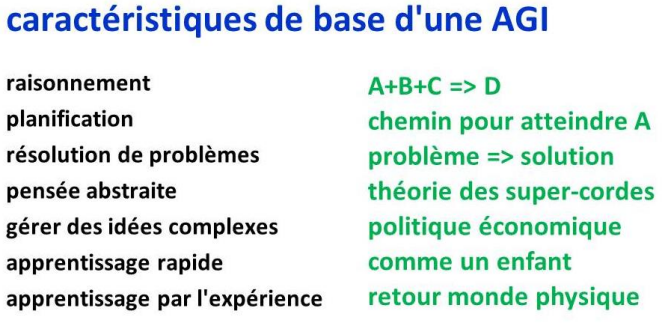
L'avènement éventuel d'une AGI dépend à la fois des progrès matériels et de notre compréhension toujours en devenir du fonctionnement du cerveau humain qui fait partie du vaste champ de la neurophysiologie, coiffant des domaines allant de la neurobiologie (pour les couches "basses") à la neuropsychologie (pour les couches "hautes"). Le fonctionnement du cerveau apparait 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 ("clock" de 100 Hz maximum).

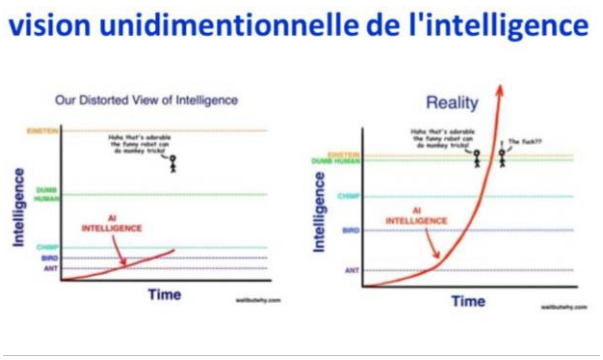
C'est aussi un engin très efficace du point de vue énergétique, ne consommant que 20W, soit l'équivalent d'un laptop équipé d'un processeur Intel Core i7.

Quelques cas pratiques d'usage de l'AGI ont été définis par le passé, comme le test de Turing (agent conversationnel 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 passer avec succès les examens, celui du chercheur capable de mener des travaux de recherche et enfin, celui du salarié de tel ou tel métier.





Les prédictions sur l'avènement de l'AGI sont souvent associées à un usage quelque peu abusif et prospectif de la loi de Moore. Elles s'appuient aussi sur une vision simpliste et unidimentionnelle de l'intelligence.



L'**Artificial Super Intelligence** (ASI) serait une intelligence largement supérieure à l'Homme. C'est la continuité logique 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 !

A vrai dire, une AGI serait d'emblée largement supérieure à l'Homme car elle accèderait facilement à tout le savoir humain déjà numérisé[70].

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. Le 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 apparaitrait très peu de temps après l'AGI, l'ordinateur faisant preuve de capacités 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 donc fonctionnent les data-centers. Mais 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 ces processus.

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 [71]!

---

[70] 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. Cf 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.

[71] 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.

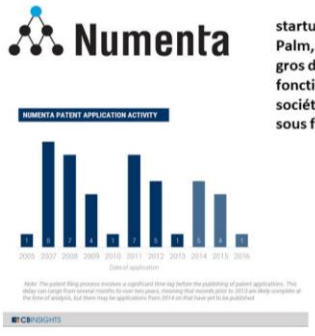
Dans la plupart des prédictions sur l'avènement de l'ASI, il est fait état de la difficulté à la contrôler. Une majeure partie des prédictions envisagent qu'elle soit même néfaste pour l'homme malgré son origine humaine. Elles évoquent une course contre la montre entre startups et grandes entreprises 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.

Dans la pratique, l'IA d'aujourd'hui va déjà bien au-delà des capacités humaines, notamment lorsque la mémoire est en jeu. La capacité des systèmes experts, et notamment d'IBM Watson, à brasser d'énormes volumes d'information fournit des capacités déjà largement inaccessibles à n'importe quel humain, même surdoué. L'ASI correspond donc à un mélange des genres entre les domaines où l'homme est déjà dépassé et ceux où il ne l'est pas encore et le deviendrait éventuellement.

Le sujet de l'AGI est en tout cas un thème de recherche très 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 de systèmes experts.

Les travaux de startup dans l'AGI sont maintenant courants. Il y a **Numenta** qui planche sur la simulation du cortex depuis plus d'une décennie avec ses réseaux de neurones HTM (Hiearchical Temporal Memory). Et puis **Kernel** qui souhaite créer une prothèse neuronale pour l'Homme[72] ! Google **DeepMind** a fait parler de lui en présentant **PathNet**, une sorte de réseau d'agents intégrant des réseaux de neurones et capable d'identifier par lui-même la meilleure combinaison de réseaux de neurones pour atteindre un objectif donné.



startup créée par Jeff Hawkins, le fondateur de Palm, auteur de l'excellent "Intelligence"
gros déposant de brevets dans la simulation du fonctionnement du cortex
société de création d'IP qui la commercialise sous forme de licences

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"

L'AGI a aussi eu son camp de détracteurs et dans toutes les décennies récentes (liste *ci-dessous*). En France, nous avons le chercheur Jean-Gabriel Ganacia[73] auteur du « Mythe de la singularité ».

---

[72] Cf https://medium.com/@bryan_johnson/kernels-quest-to-enhance-human-intelligence-7da5e16fa16c#.i0pveroe9.

[73] Cf http://internetactu.blog.lemonde.fr/2017/06/25/la-singularite-ca-ne-tient-pas-la-route/

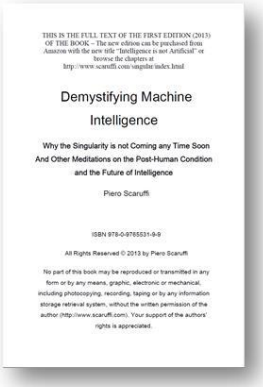
J. R. Lucas, Minds, Machines and Gödel, Philosophy XXXVI, 1961.

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

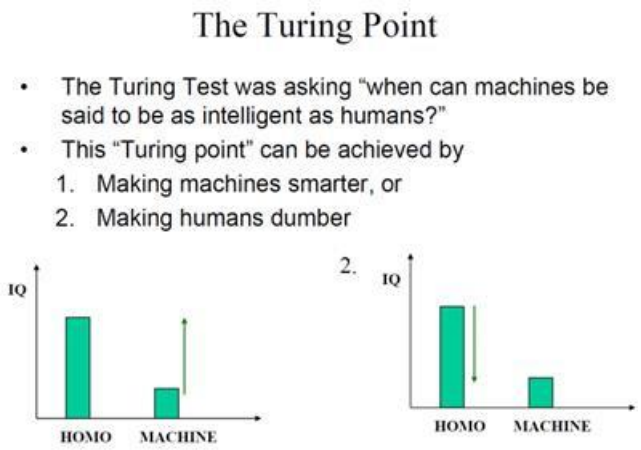
R. Penrose, "The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics", 1989.

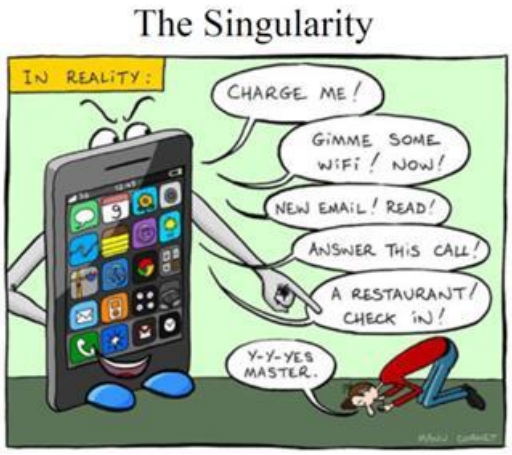
Nills Nilsson, "Human-Level Artificial Intelligence? Be Serious!", 2005.

D. Chalmers, Contemporary Philosophy of Mind: An Annotated Bibliography, Part 4: Philosophy of Artificial Intelligence.



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. C'est la thèse ironique de **Piero Scaruffi**[74] et aussi celle de **Nicholas Carr**[75].





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.

Imiter ou s'inspirer du cerveau humain

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.

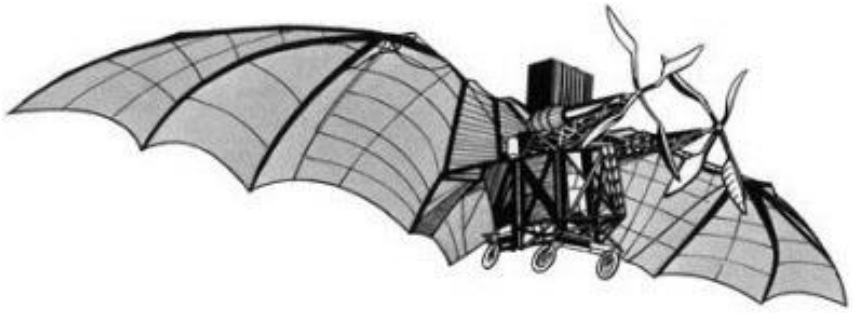
C'est un peu comme la magie. Tant que l'on ne connait pas le truc, c'est de la magie voire de l'art. Une fois qu'on le connait, c'est une technique, souvent très simple, si ce n'est évidente.

[74] Dans Artificial Intelligence and the Singularity, octobre 2014.

[75] Dans Is Google making us stupid, dans The Atlantic, juillet 2008.

L'intelligence humaine est un peu du même ressort quand on n'en connait pas le fonctionnement exact. Elle préserve ce côté mystérieux et inimitable, presque immatériel, comme une âme qui n'aurait pas d'existence physique.

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 !



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 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-dessous*), dont les premiers vols ont eu lieu en 1969 (*ci-dessus à droite*).

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. Un aigle se pose en quelques secondes et presque n'importe où ! C'est la puissance 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, le machine 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'information 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. 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 pas tâtonnements, par intégration de briques algorithmiques et logicielles disparates, et pas seulement via la force brute de la machine.

Décoder le cerveau

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 connaitre son fonctionnement pour découvrir des traitements de certaines pathologies neurodégénératives.

De nombreuses initiatives de recherche nationales et internationales ont été lancées dans ce sens. **Inventoriées ici**, 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** vise à simuler numériquement le fonctionnement d'un cerveau. Lancé après la réponse à un appel d'offre par Henry Markram de l'EPFL de Lausanne, un chercheur à l'origine du **Blue Brain Project** lancé en 2005, qui vise à créer un cerveau synthétique de mammifère. Construit à partir d'un supercalculateur Blue Gene d'IBM et faisant tourner le logiciel de réseau de neurones de Michael Hines, le projet vise à simuler de manière aussi réaliste que possible des neurones[76].

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. Il est critiqué **ici** et là. 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 !

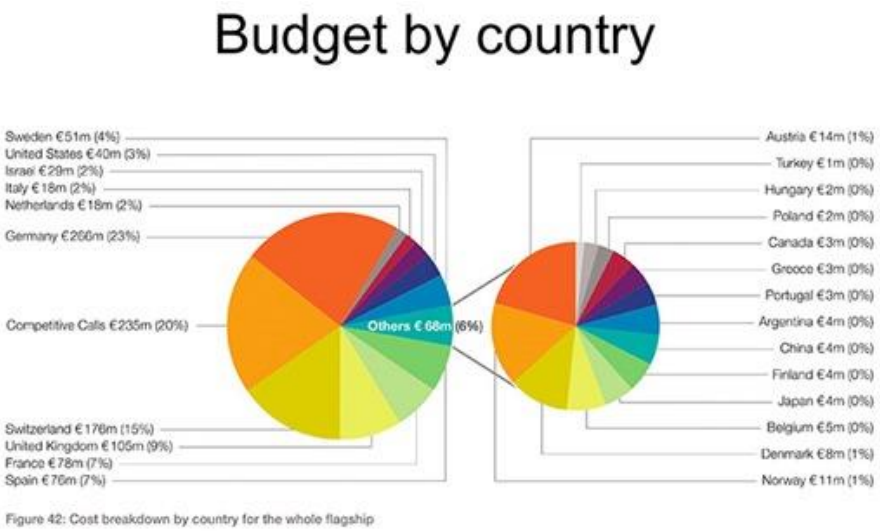


Figure 42: Cost breakdown by country for the whole flagship

[76] J'ai eu l'occasion de rentrer un peu plus en détails dans ce projet dans http://www.oezratty.net/wordpress/2017/startups-bidouille-cerveau-autres/ publié en juin 2017.

Dans la pratique, 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 du fonctionnement des neurones dans le cerveau.

Les USA ne sont pas en reste avec la **BRAIN Initiative** annoncée par Barack Obama en 2013. Elle vise 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.

Le budget annuel est 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 nanotechnologies pour mesurer l'activité individuelle de cellules nerveuses, à commencer par celles des mouches drosophiles.

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 (*exemple ci-dessus* avec les principales liaisons nerveuses internes au 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.

Reste aussi, côté neurobiologie, à comprendre le processus d''apprentissage 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 ? 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.

Copie du cerveau

Dans "The Singularity is Near[77]", Ray Kurweil fantasme 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 !

[77] Téléchargeable librement ici : http://stargate.inf.elte.hu/~seci/fun/Kurzweil,%20Ray%20-%20Singularity%20Is%20Near,%20The%20(hardback%20ed)%20%5Bv1.3%5D.pdf.

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[78].

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 !

Quoi qu'il en soit, l'information est 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). Et il n'y a que quelques nanomètres entre les dendrites et les terminaisons des axones !

La communication entre les deux 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 elles 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 ?

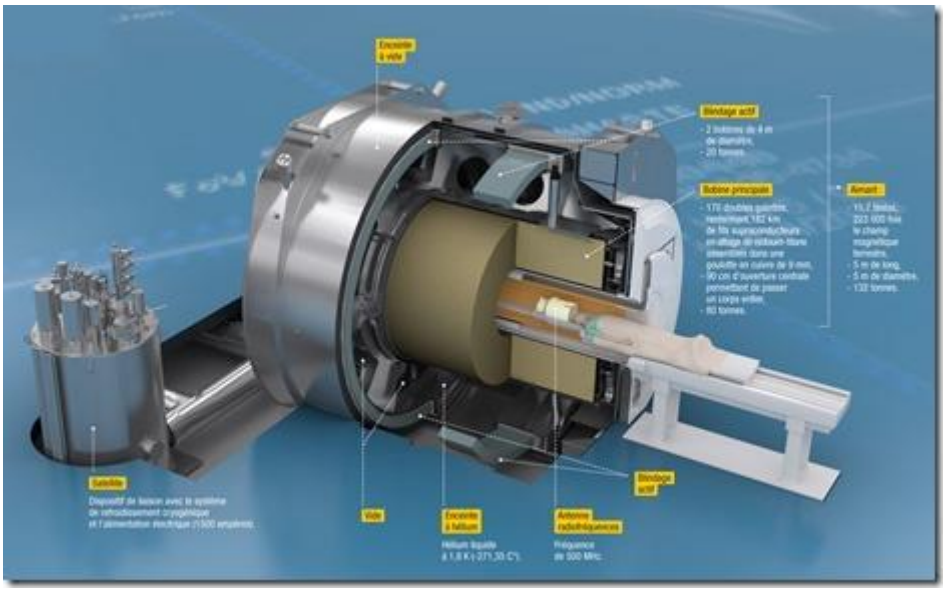
Il n'y a pas 36 solutions : il faut passer par des ondes électromagnétiques, et avec une précision de l'échelle du nanomètre. Aujourd'hui les scanners utilisent généralement trois technologiques : la tomodensitographie qui mesure la densité de la matière par rayons X, les PET scanners qui détectent des traceurs biologiques radioactifs par émission de photons et l'IRM qui détecte les corps mous par résonnance magnétique

[78] Découverte confirmée par des chercheurs du MIT début 2016. Cf http://www.extremetech.com/extreme/123485-mit-discovers-the-location-of-memories-individual-neurons.

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 fait en tout cas bien avancer l'état de l'art. 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, et dont le 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[79].

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 est prévue pour 2018, 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 a court : plus on veut observer l'infiniment petit, plus l'instrument est grand et cher. Comme pour les accélérateurs de particule et le LHC 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.

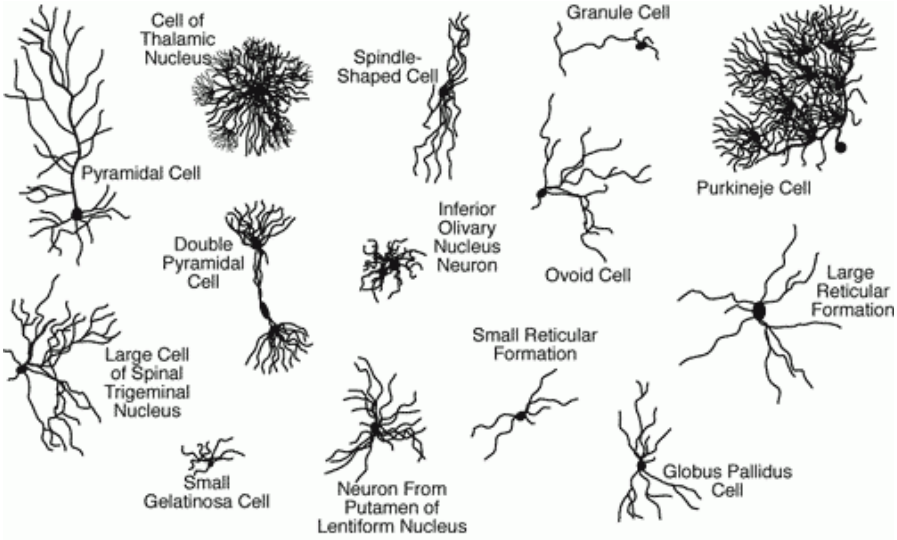
[79] L'aimant a été conçu avec le concours du CEA-Irfu, l'Institut de Recherche sur les lois Fondamentales de l'Univers, qui a réutilisé ses acquis issus de la création des aimants supraconducteurs du Large Hadrons Collider du CERN de Genève. Il est fabriqué par Alstom-GE à Belfort, l'intégration du 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.

D'où l'intérêt de la solution légère et, en apparence, très élégante, de OpnWatr évoquée dans un autre post, mais qui n'a pas encore fait ses preuves. Elle pourrait sortir du gué dès 2018, au même moment qu'Iseult. La confrontation sera plus qu'intéressante !

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. 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 connaitre avec précision la position 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 reconnaitre.

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[80].



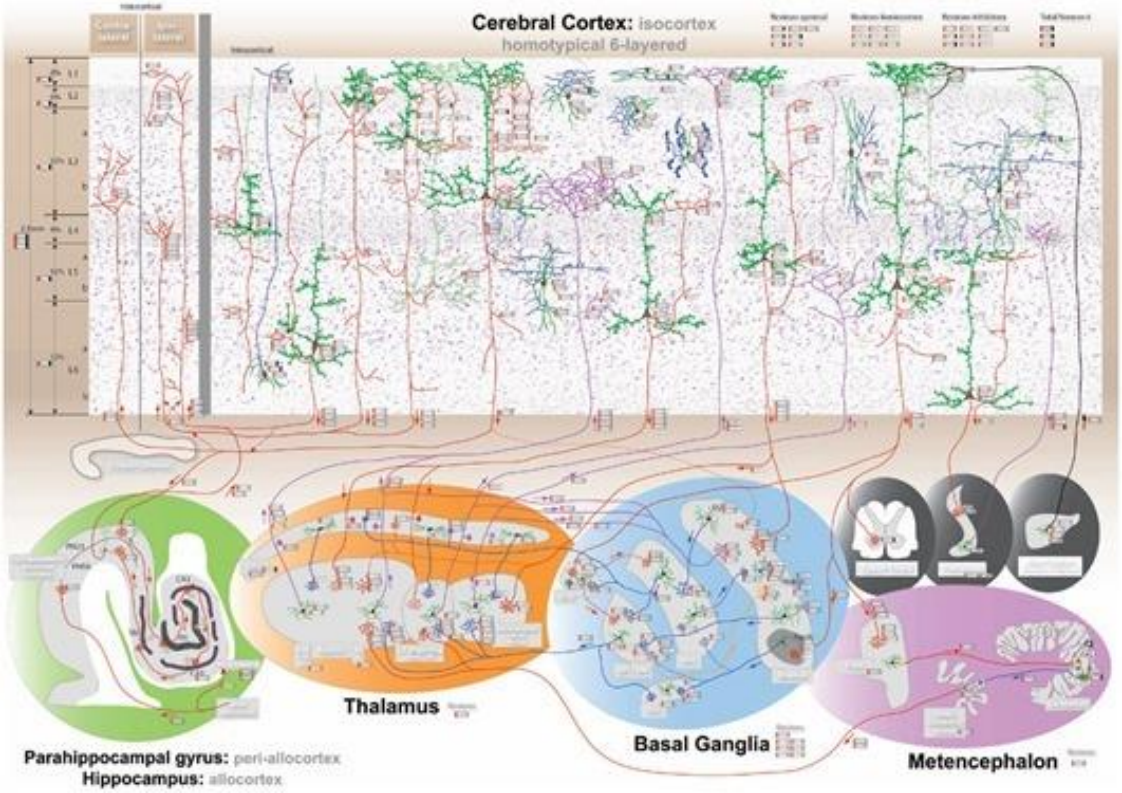
(source du schéma ci-dessus sur quelques exemples de neurones cérébrales : **http://neuromorpho.org**)

Mais poursuivons dans la découverte fascinante de la complexité du cerveau. Celui-ci contient plusieurs centaines de types de neurones différents (**source**), les illustrations précédente et suivante 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, qui contrôlent les mouvements appris.

[80] 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 les dérives des exponentielles.

Cette complexité se retrouve aussi au niveau moléculaire avec de nombreuses protéines et hormones intervenant dans la transmission d'influx neuronaux, comme décrit dans **Deep Molecular Diversity of Mammalian Synapses: Why It Matters and How to Measure It**. Parmi les 20 000 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 ! C'est dire la richesse de la soupe de protéines qui gouverne le cerveau, dont l'actine qui structure la forme mouvante des neurones !

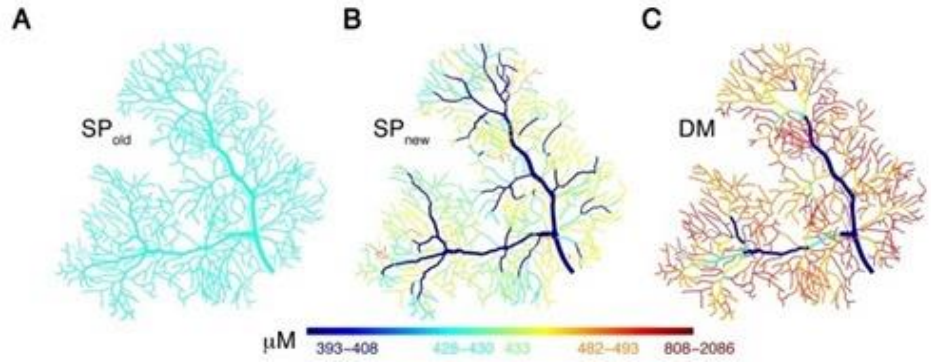
Le cerveau d'un fœtus comprendrait plus de mille milliards de neurones, qui meurent rapidement. On perd en fait des neurones dès sa naissance, comme si une matrice s'évidait pour prendre forme progressivement au gré des apprentissages. Le cerveau d'un enfant comprendrait plus de 100 milliards de neurones, et plus de 15 trillions de synapses et 150 milliards de dendrites.



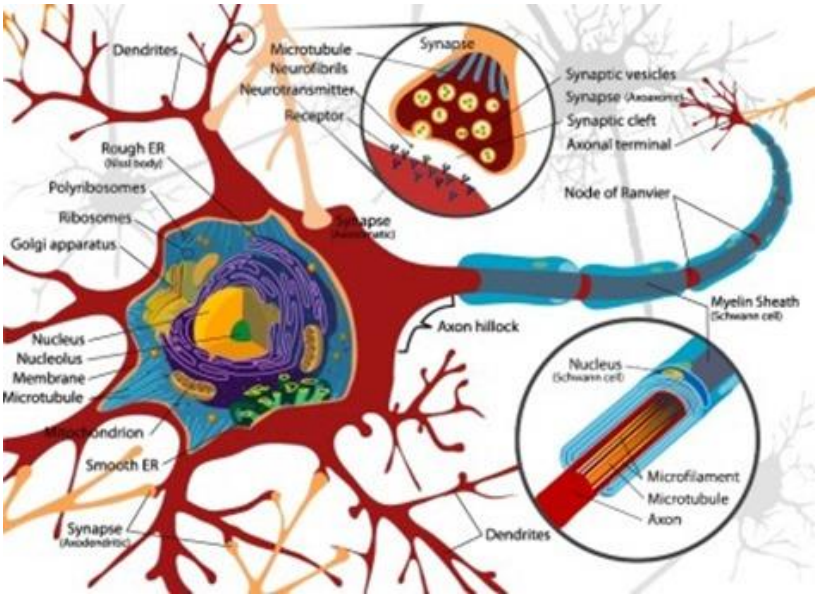
(source de l'illustration ci-dessus)

Un cerveau adulte comprend environ 85 milliards de neurones dont 16 milliards dans le cortex et environ 56 milliards dans le cervelet, 10 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). Il consomme environ 20 Watts 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 !



Plus de la moitié des neurones du cerveau sont situées dans le cervelet. Il gère les automatismes appris comme la marche, la préhension, les sports, la conduite, le vélo, la danse ou la maitrise des instruments de musique. Un neurone du cervelet contient environ 25 000 synapses le reliant aux terminaisons d'axones d'autres neurones. Ceux du cortex qui gèrent les sens et l'intelligence comprennent chacun de 5000 et 15 000 synapses.



(**source** du schéma qui l'explique très bien)

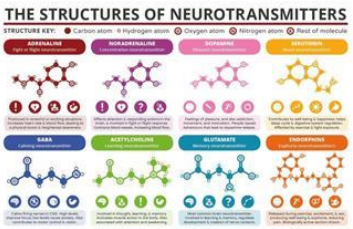
Le cerveau est aussi rempli de cellules gliales qui alimentent les neurones et en contrôlent le fonctionnement via la myéline qui entoure les axones et divers autres mécanismes de régulation. Il y en a au moins autant que de neurones dans le cerveau, ce qui ajoute un niveau de complexité de plus.

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. 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.

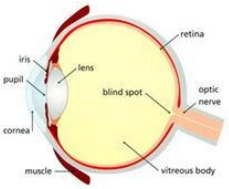
Le cerveau est imbattable dans sa densité, sa compacité et son parallélisme. Par contre, les ordinateurs nous dépassent dans leur capacité de stockage et de traitement de gros volumes de données. Si l'on aura bien longtemps du mal à scanner un cerveau au niveau des neurones, il n'en reste pas moins possible d'en comprendre le fonctionnement par tâtonnements.

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.

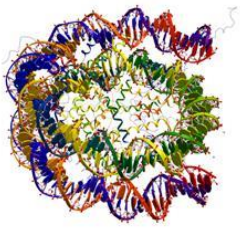
86 milliards de neurones
600 trillions de synapses



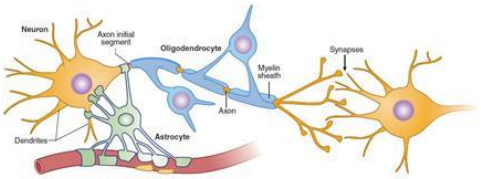
8 neurotransmetteurs différents



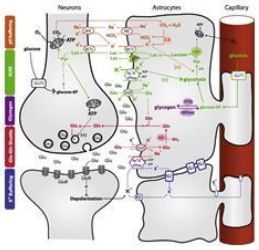
6,5 millions de cônes
90 millions de bâtonnets
1 million d'axones



6000 gênes spécifiques



88 milliards de cellules gliales dans le cerveau



des interactions complexes

Mais il n'est pas nécessaire de maitriser 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 maitriser 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 !

Placer l'intelligence de la machine dans la prolongation de celle de l'homme et sur une simple courbe exponentielle n'a pas beaucoup de sens, comme dans **The AI Revolution: Our Immortality or Extinction** de Tim Urban.

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. L'intelligence artificielle sera supérieure à celle de l'homme dans de nombreux domaines et pas dans d'autres, comme aujourd'hui.

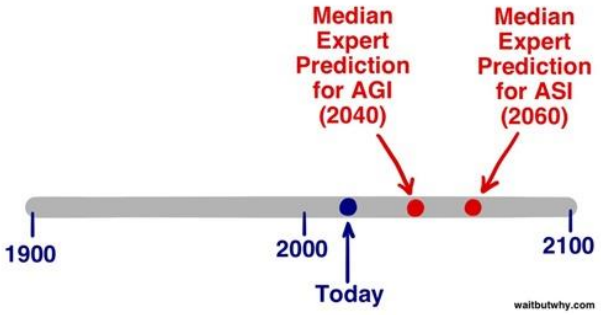
Elle sera simplement différente et complémentaire. Tout du moins, à une échéance raisonnable de quelques décennies.

Vu du versant de l'optimisme et couplée à la copie du cerveau dans le silicium, l'ASI aurait un impact indirect : l'immortalité de l'homme, conséquence des découvertes générées par l'ASI.

C'est évidemment faire abstraction de ce qui ne peut pas encore se faire de manière entièrement numérique. Les progrès dans la santé sont contingentés par l'expérimentation qui se fait encore in-vivo et in-vitro.

L'expérimentation in-silico – de manière entièrement virtuelle et numérique – des processus biologiques est un domaine en plein devenir. Il se heurte pour l'instant à des obstacles proches de l'insurmontable, même en intégrant les merveilles des exponentielles de progrès et de la loi de Moore. La recherche scientifique dans la santé en est donc toujours réduite à mener des expérimentations itératives et plutôt lentes, même avec les appareillages les plus modernes. Avec ou sans IA, cela reste immuable.

D'ailleurs, les meilleures solutions d'IA comme l'usage d'IBM Watson dans la cancérologie s'appuient sur le corpus issu de toutes ces expérimentations. Il a une base physique et réelle. On pourra certainement automatiser l'expérimentation biologique encore plus qu'aujourd'hui dans la recherche de thérapeutiques, mais cela restera toujours du domaine du biologique, pas du numérique, donc plutôt lent et pas très scalable.



On arriverait au stade de l'AGI entre 2030 et 2100 selon les prévisions, et de l'ASI quelques décennies après. On se demande d'ailleurs ce qui expliquerait le délai entre les deux au vu du facteur d'accélération lié au matériel.

IA et infrastructure informatique

Nous avons étudié dans la partie précédente les évolutions de l'IA côté méthodes, outils et logiciels. Passons ici à son substrat matériel, fait d'ordinateurs, de processeurs plus ou moins spécialisés, de systèmes de stockage et de capteurs. Leurs évolutions respectives contribuent aussi à améliorer la qualité des solutions d'intelligence artificielle.

L'un des domaines les plus importants pour l'IA sont les serveurs d'entrainement pour le deep learning. En effet, si celui-ci donne de très bons résultats, comme dans la reconnaissance d'images, il est très consommateur de ressources dans sa phase d'entrainement. Il faut facilement 1000 à 10000 fois plus de puissance machine pour entrainer un modèle de reconnaissance d'images que pour l'exécuter ensuite. Cela explique pourquoi, par exemple, les GPU et TPU font environ 100 Tflops/s tandis que les briques neuronales des derniers Kirin 970 de Huawei et de l'A11 Bionic ne font que 1 à 2 Tflops/s.

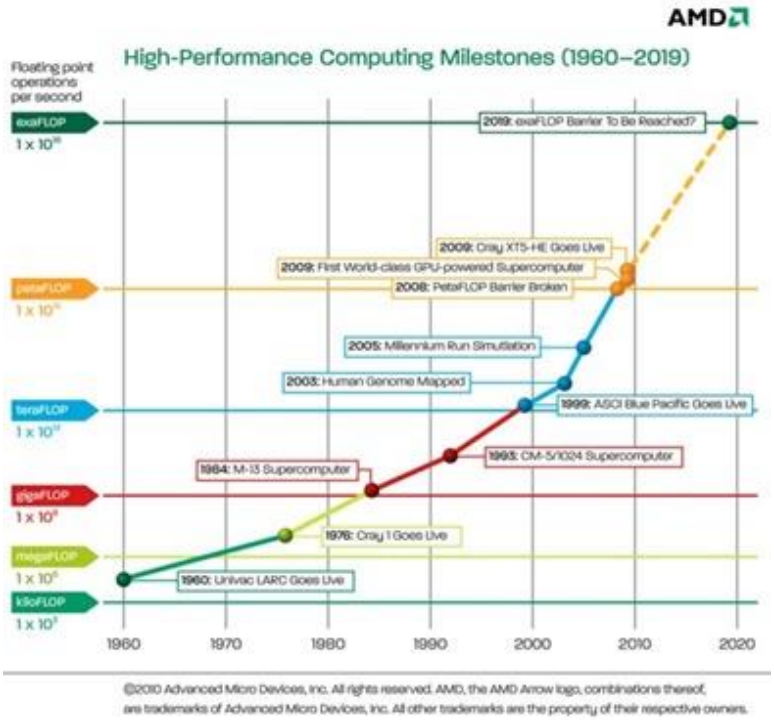
Processeurs

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.

Cela peut concerner 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 entraine 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.

Poursuivre la loi de Moore

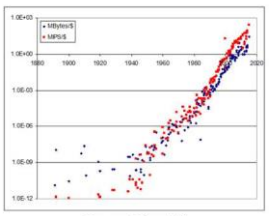
La loi de Moore est censée s'appliquer à des solutions commercialement disponibles, et si possible, en volume. Or ce n'est pas toujours le cas. Ainsi, l'évolution de la puissance des supercalculateurs est mise en avant comme un progrès technique validant la loi de Moore. Or, ces calculateurs sont créés 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 (aérospatial, génomique, météo). Ce que l'on peut observer dans la belle exponentielle ci-dessous issue d'AMD.



Dans la plupart des cas, ces technologies “de luxe” sont intégrées dans des produits grand public après quelques années. 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.

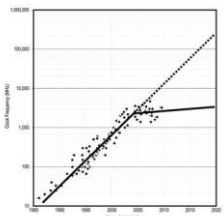
Dans la pratique, le nombre de transistors continue d’augmenter régulièrement mais pas la fréquence d’horloge, d’où le choix des architectures multicoeurs. Par ailleurs, il existe une barrière méconnue, celle de Landauer, qui définit le niveau minimum d’énergie nécessaire pour gérer un transistor. Et elle pourrait être atteinte d’ici 2035.

transistors / puces

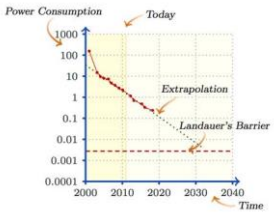


Mo et Mips / $

source : http://nerdfever.com/moores-law-isnt-slowing-down/



barrière de la chaleur
limite la fréquence d’horloge des processeurs



barrière de Landauer
limite basse de la consommation d’énergie d’un transistor

La question est revenue au-devant de la scène alors que cette loi fêtait ses 50 ans d’existence. Un anniversaire commenté pour annoncer la fin de ses effets, tout du moins dans le silicium et les technologies CMOS. Cette technologie est sur le point d’atteindre un taquet aux alentours de 5 nm d’intégration sachant que l’on atteint maintenant les 10 nm.

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.

L'excellent dossier **After Moore's Law**, paru dans The Economist en mars 2016, détaille bien la question en expliquant pourquoi la loi de Moore des transistors CMOS pourrait s'arrêter d'ici une douzaine d'année lorsque l'on descendra au niveau des 5 nm d'intégration. Et encore, la messe n'est pas encore dite. A chaque nouvelle génération d'intégration, les fondeurs se demandent s'ils vont pouvoir faire descendre réellement le cout de fabrication des transistors. En-dessous de 14 nm, ce n'est pas du tout évident. Qui plus est, Intel met plutôt l'accent sur la consommation d'énergie que sur les performances brutes. Ainsi, un Core i7 7500U génération Kaby Lake en 14 nm pour laptop n'a-t-il aujourd'hui que deux cœurs et tourne à environ 3 GHz !

Alors, la loi de Moore est foutue ? Pas si vite ! Il faudra tout de même trouver autre chose, et en suivant divers chemins de traverse différents des processeurs en technologie CMOS. Elle avance par hoquets. Il reste encore beaucoup de mou sous la pédale pour faire avancer la puissance du matériel et sur lequel l'IA pourrait surfer.

En 2015, IBM et Global Foundries créaient une première en testant la création d'un processeur en **technologie 7 nm** à base de silicium et de germanium, battant le record d'Intel qui est descendu en production à 10 nm tout comme Samsung et TSMC. L'enjeu clé est de descendre en intégration sans que les prix n'explosent. Or, la gravure en extrême ultra-violet qui est nécessaire pour "dessiner" les transistors sur le silicium est complexe à mettre au point et plutôt chère.

trigate transistors
less leaks, power saving
more complex steps

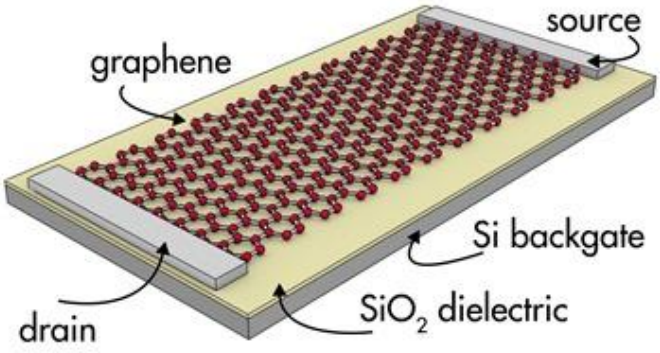
EUV (extreme ultra violet) etching
drawing finer lines
harder than planned to implement

double and triple patterning
avoids small lines overlaps
but more process steps

Le multi-patterning, que j'explique **ici**, permet d'en contourner les limitations. Mais il coute tout aussi cher car il ajoute de nombreuses étapes à la fabrication des chipsets et peut augmenter le taux de rebus. La loi de Moore s'exprime en densité de transistors et aussi en prix par transistors. Si la densité augmente mais que le prix par transistor augmente aussi, cela ne colle pas pour les applications les plus courantes.

Nouveaux transistors

Cela permettrait d'accélérer leur **vitesse de commutation** et augmenter grâce à cela la fréquence d'horloge des processeurs. Cela peut passer par exemple par des portes au graphène IBM avait **annoncé** en 2011 avoir produit des transistors au graphène capables d'atteindre une fréquence de 155 GHz, et en 40 nm.



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.

GPU

Les GPU constituent la solution matérielle la plus largement déployée pour accélérer les réseaux de neurones et le deep learning. Le leader de ce marché est l'américain Nvidia qui fournit des chipsets et des cartes équipés de GPU dépassant les 5000 cœurs. 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[81].

Nvidia avait présenté au CES 2016 de Las Vegas sa carte PX2 pour l'automobile qui intègre deux processeurs X1 comprenant 256 GPU. Les GPU Nvidia sont utilisés pour simuler des réseaux de neurones. Cette carte doit être réfrigérée par eau car elle consomme plus de 200 W.

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.

Cette dernière 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[82] ! On les trouve dans les cartes Tesla V100 qui équipent notamment les serveurs DGX-1 par paquet de 8, totalisant 40 960 cœurs pour $150K ainsi que dans l'architecture de référence HGX-1 destinée aux supercalculateurs hyperscale comme Microsoft Olympus et Facebook Big Basin. Nvidia propose aussi une version station de travail de la DGX-1, dotée de quatre cartes V100.

---

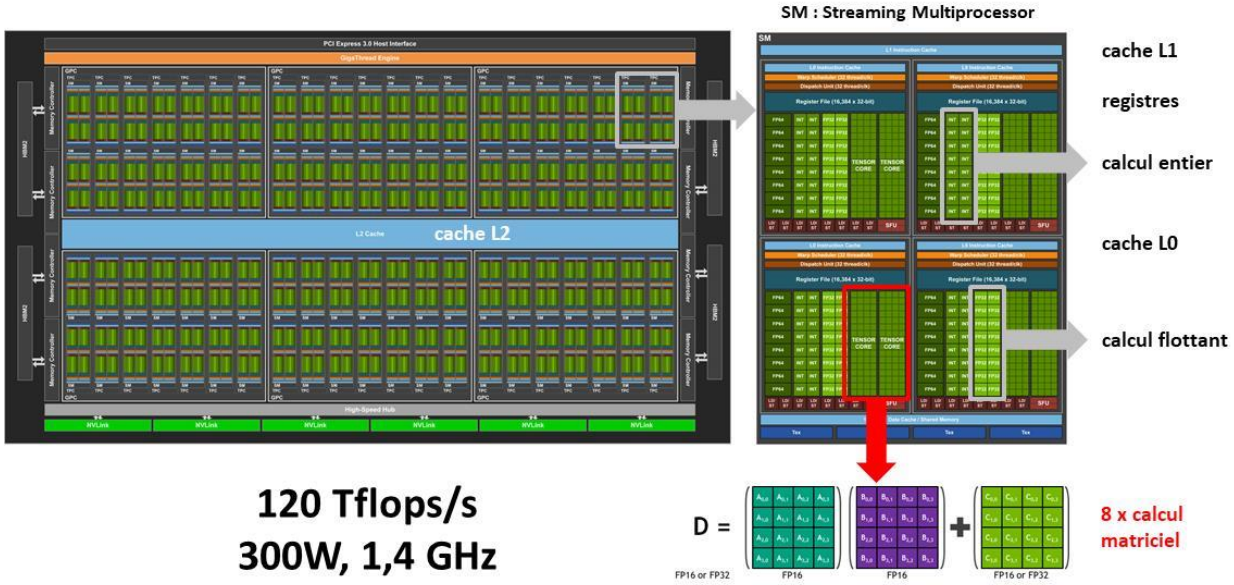
[81] Cf AMD Too Late To NVIDIA Deep Learning Party, de Boris Suvorov, décembre 2016. Et la dernière génération de cartes AMD, lancées en juin 2017, les Radeon Instinct M125, a une puissance théorique de 24,6 Tflops, à comparer aux dernières Nvidia V100 qui atteignent 120 Tflops.

[82] Voir pas mal de détails ici : http://www.anandtech.com/show/11367/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.

Le français **Qwant** a annoncé en 2017 faire l'acquisition d'une douzaine de ces serveurs lors d'un partenariat avec Nvidia pour les utiliser dans son moteur de recherche. Idem pour **OVH** qui est également en train de les déployer.

Jusqu'à présent, les GPU comprenaient une myriade de cœurs à même d'effectuer des opérations mathématiques simples (multiplications, divisions, additions, soustractions). 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.

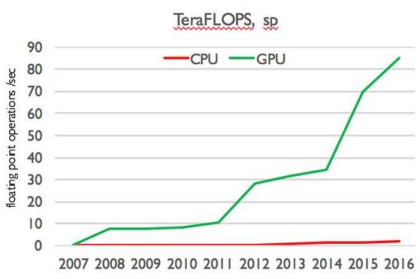
Ces GPU comprennent 80 « streaming multiprocessors », comprenant un total de 5120 cœurs CUDA traditionnels (avec 64 cœurs en flottant 32 bits, 32 cœurs flottant 64 bits et 64 cœurs entier par SM) et 640 « tensor cores » (8 par SM). Cette architecture présente l'avantage d'être assez flexible et générique et de s'adapter à de nombreux types de traitements. Elle est par ailleurs très bien supportée côté logiciels et frameworks.

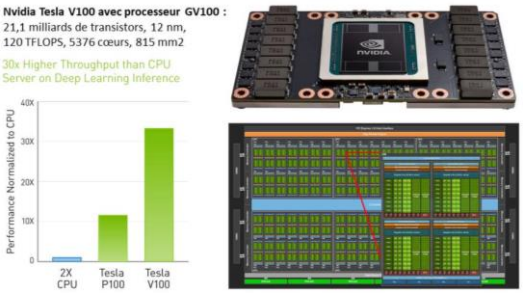
Nvidia DGX-1 : 8xTesla V100
960 TFLOPS, 40 960 cœurs, 5120 coeurs Tensor, $150K

Nvidia Titan Xp
12 TFLOPS, 3840 cœur, $1,3K

Nvidia Drive PX-2
8 TFLOPS

Intel Core i7
< 0,2 TFLOPS



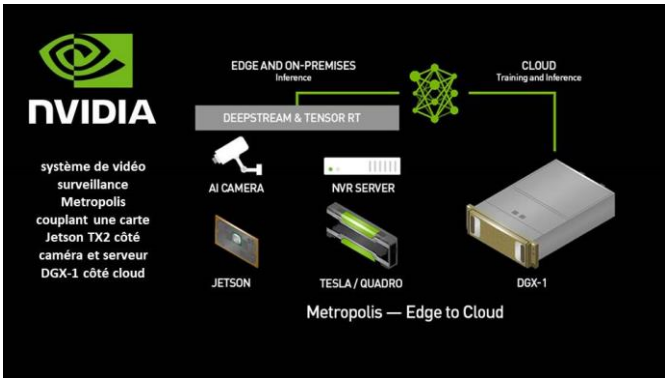


Nvidia propose aussi des stations de travail équipées également de GPU puissants ainsi que des cartes Jetson pour l'informatique embarquée. L'architecture est complète, des objets connectés aux serveurs, permettant de répartir les traitements en fonction de leur nature ainsi que des réseaux de télécommunications utilisés.

Nvidia DIGITS DevBox

4 TITAN X GPUs with 12GB per GPU
64GB DDR4
Asus X99-E WS with 4-way PCI-E Gen3 x16
Core i7-5930K 6 Core 3.5GHz desktop CPU
3 x 3TB SATA 6Gb 3.5" HDD in RAID5
512 GB PCI-E M.2 SSD cache for RAID
250 GB SATA 6Gb Internal SSD
1600 W Power Supply
NVIDIA CUDA Toolkit, DIGITS SW, cuDNN
Ubuntu 14.04
Caffe, Theano, Torch, BIDMach

Asus Server with Nvidia VCA



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.

Enfin, 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, soit 10 fois le niveau de performance de la génération précédente, les Drive PX 2. La Pegasus exploite quatre processeurs embarqués dont deux de la série Xavier embarquant un GPU de la série Volta. 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 entrainés sur des serveurs Nvidia DGX-1.

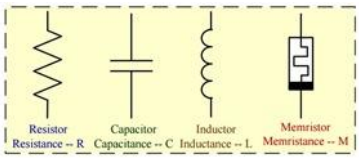
Les GPU de Nvidia s'améliorent avec leurs tensors et sont conçus pour supporter des charges de travail très variées pour toutes sortes d'applications du domaine de l'IA. A force de persévérance, d'innovation continue et d'une approche « plate-forme » soutenue, Nvidia est en train de devenir le leader de-facto des processeurs dédiés à l'IA.

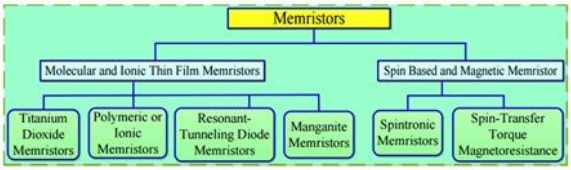
Memristors

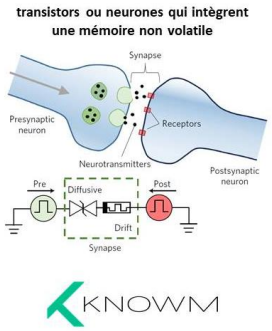
Les memristors ont fait son apparition en 2008 chez HP après avoir été conceptualisée en 1971 par le sino-américain **Leon Ong Chua**. Ce sont des composants électroniques capables de mémoriser un état en faisant varier leur résistance électrique par l'application d'une tension. Un peu comme les cristaux liquides bistables qui servent dans (feu) les liseuses électroniques. La valeur modifiable de la résistance permet de stocker de l'information.

Les memristors peuvent aussi être intégrés au côté de composants actifs classiques dans des unités de traitement. C'est très bien expliqué dans **Memristor: From Basics to Deployment** de **Saraju Mohanty**, publié en 2013, d'où sont extraits les deux schémas ci-dessous. Le second présente les différents types de memristors actuellement explorés.

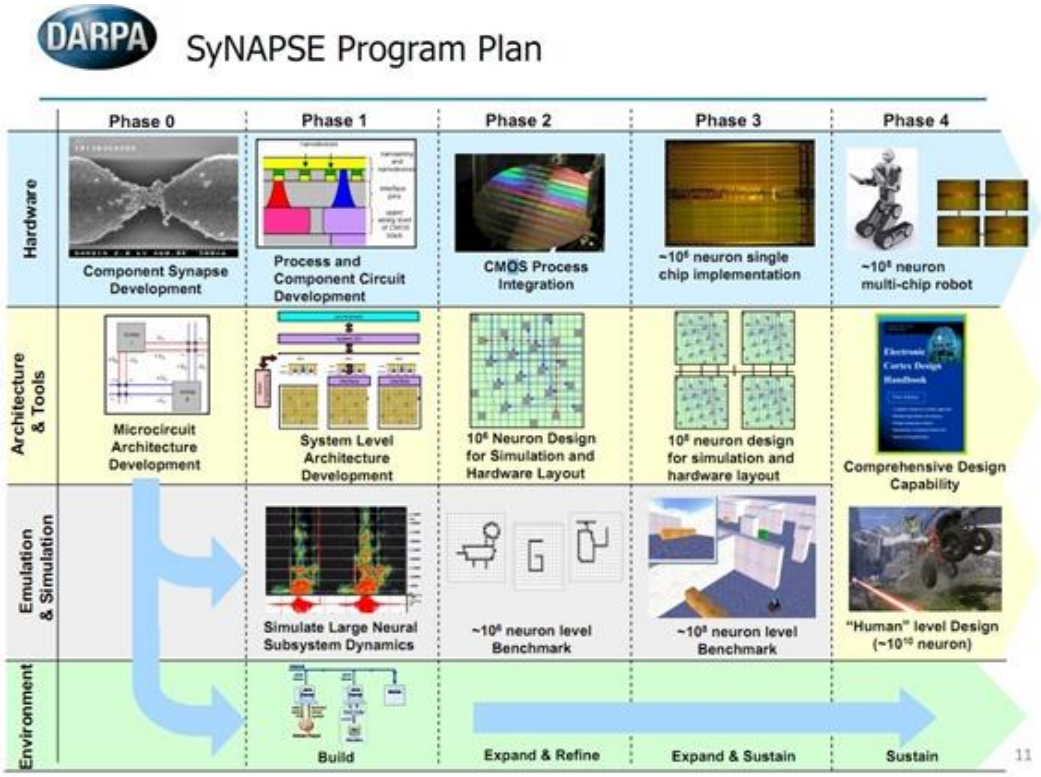
Ces composants sont intégrables dans des puces au silicium utilisant des procédés de fabrication plus ou moins traditionnels (**nanoimprint lithography**), en ajoutant une bonne douzaine d'étapes dans la production, et avec des matériaux rares comme les oxydes de titane.







Les memristors ont été développés dans le cadre des projets de recherche du programme **SyNAPSE** de la DARPA. **HP** a été le premier à en prototyper en 2008, avec de l'oxyde de titane. Il en existe de plusieurs types, pouvant généralement être fabriqués dans les lignes de productions de chipsets CMOS traditionnelles, mais avec des procédés spécifiques de dépôt sous vide de couches minces de matériaux semi-conducteurs.



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** ont cependant annoncé 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.

D'autres laboratoires de recherche et industriels planchent aussi sur les memristores et les réseaux de neurones matériels :

- **IBM** planche avec l'**ETH** de Zurich (le CNRS suisse) sur des ordinateurs à base de memristors. Ce même ETH développe un **memristor** capable de stocker trois états à base de pérovskite (titanate de calcium) de 5 nm d'épaisseur. Cela pourrait servir à gérer de la logique floue.
- Des chercheurs de l'Université Technologique du Michigan ont **annoncé début 2016** avoir créé des memristors à base de bisulfite de molybdène qui ont un comportement plus linéaire.
- Des **chercheurs du MIT** ont annoncé 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) et s'appuie sur des fils microscopiques pour connecter plus efficacement des cœurs et neurones dans des circuits spécialisés.
- Enfin, la start-up californienne **Knowm** a lancé le **premier 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.
- L'ANR française a financé le projet collaboratif **MHANN** associant l'INRIA, l'IMS de Bordeaux et Thalès pour créer des memristors ferriques. Le projet devait être terminé en 2013 et avait bénéficié d'une enveloppe de 740 K€. Difficile de savoir ce qu'il en est advenu en ligne.
- Le **CNRS** en partenariat planche aussi avec Thalès sur une technologie de memristors, à l'état de recherche pour l'instant[83].

[83] Cf http://www.zdnet.fr/actualites/intelligence-artificielle-creation-d-une-synapse-artificielle-39850840.htm.

Processeurs neuromorphiques

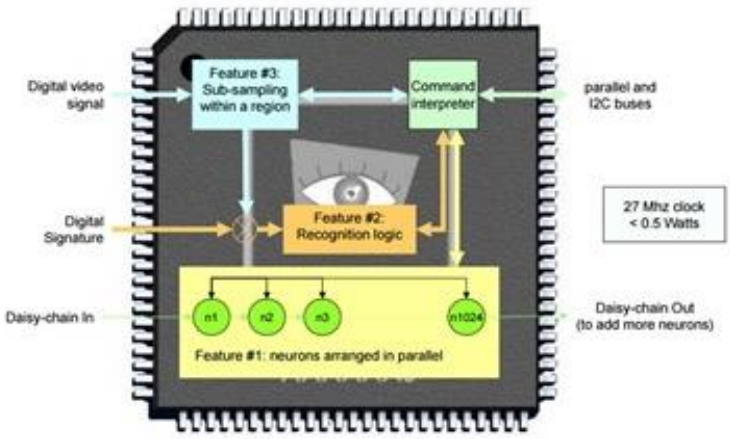
Les processeurs neuromorphiques auraient été conceptualisés pour la première fois en 1990, par Carver Mead[84].

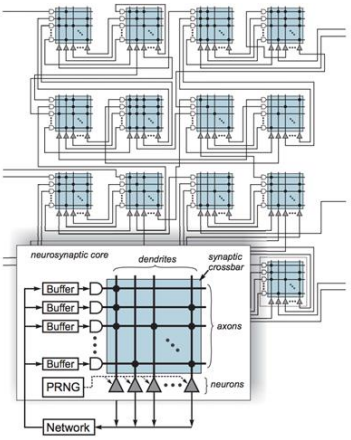
Ils sont spécialisés dans les traitements de machine learning et deep learning. Ils exécutent en parallèle les processus d'entrainement puis d'exécution d'applications de deep learning et de machine learning qui s'appuient sur des réseaux de neurones.

Ces processeurs peuvent comprendre en général des unités spécialisées dans la multiplication de matrices qui sont utilisées dans les premières couches des réseaux de neurones convolutionnels. Les dernières couches « fully connected » utilisent également des matrices qui relient les neurones d'entrées avec les neurones des couches suivantes, via des grilles comprenant les poids des synapses (*exemple ci-dessous à droite*). Mathématiquement, ce sont des unités de traitement qui multiplient des vecteurs à une dimension par des matrices pour générer des vecteurs.

Ils sont 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'entrainement[85].

Comme les GPU, ces processeurs neuromorphiques servent à gérer des modèles de réseaux de neurones de plus en plus grands, d'accélérer les phases d'entrainement, et à réduire la consommation énergétiques des machines aussi bien côté data centers que dans l'embarqué.





On peut classifier en trois catégories ces processeurs :

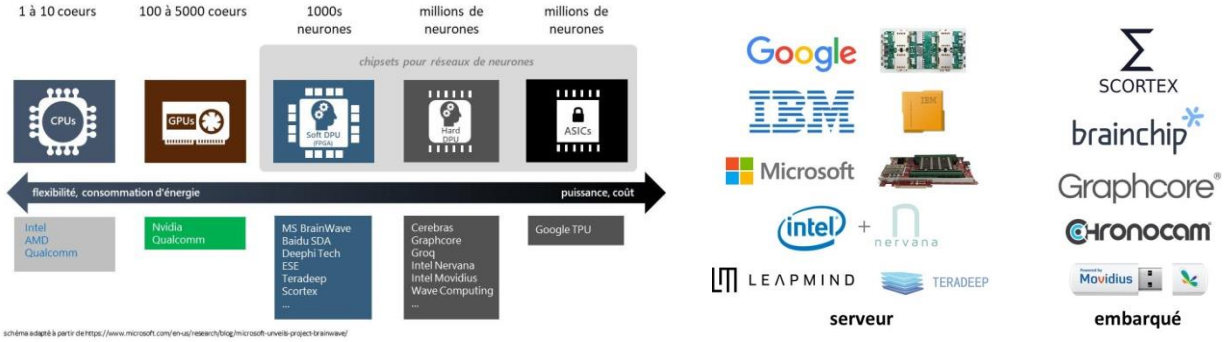
- **FPGA** : ce sont des processeurs programmables qui sont utilisés lorsque les volumes sont faibles. Ce sont des circuits dont on peut activer par logiciel les « portes » pour créer des neurones sur mesure. Ils sont un peu l'équivalent de l'impression 3D pour les chipsets : intéressants pour les faibles volumes et le prototypage rapide. C'est la technologie retenue par **Microsoft** pour ses chipsets

[84] Cf Neuromorphic Electronic Systems, Carver Mead, 1990.

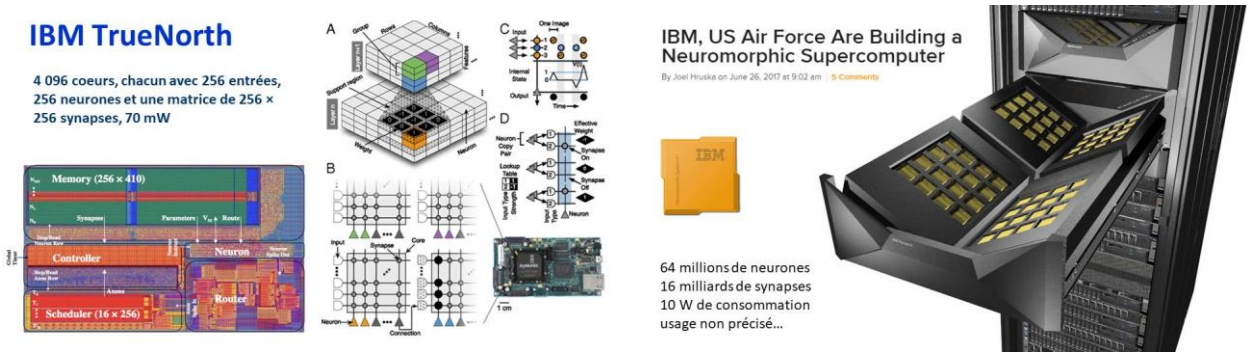
[85] C'est une approche qui est aussi adoptée par la startup grenobloise **UpMem** (2015, $3,6M) qui conçoit des circuits de traitement intégrant mémoire et calcul (Processing-In-Memory ou PIM), mais dédiés au big data. Visiblement, l'architecture semble plus proche de celle des GPU que des processeurs neuromorphiques.

Brainwave. On en trouve aussi chez diverses startups comme **Teradeep** (2014) ou le japonais **Leapmind** (2012, $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 ou les processeurs mobiles courants. Elle est adaptée aux gros volumes. 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[86].
- **Memristors** : que nous avons examiné juste au-dessus. Ce sont des circuits de réseaux de neurones qui mémorisent de manière non volatile les poids des synapses 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 programme SyNAPSE de la DARPA a aboutit en 2014 à la création par IBM de ses processeurs neuronaux **TrueNorth** 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. TrueNorth utilise des neurones impulsionnels (spiking neurons) imitent mieux le fonctionnement des neurones biologiques qui émettent des impulsions à intervalle régulier. Ces neurones semblent plutôt adaptés au traitement du langage qu'à celui des images.



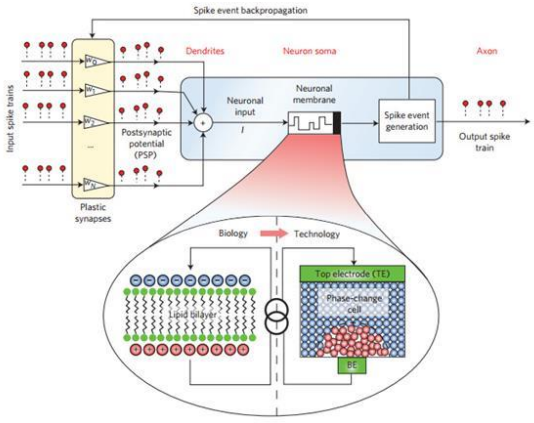
[86] L'approche de Google est décrite en détails ici : https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture/.

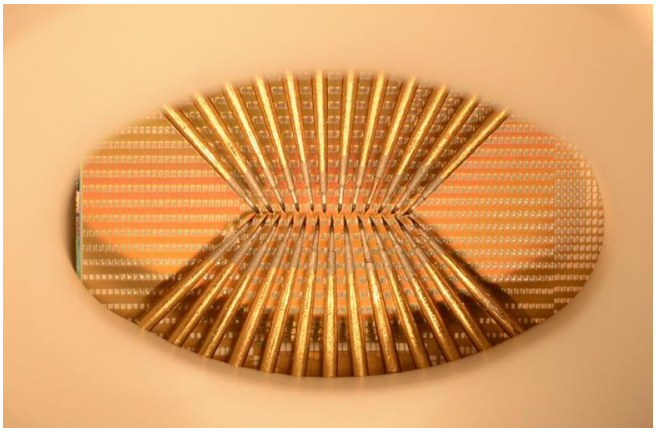
Le chipset a été fabriqué par Samsung en technologie CMOS 28 nm et avec une couche d'isolation SOI (issue du français SOITEC !) permettant de diminuer la consommation électrique et d'accélérer les traitements. Le chipsets comprend 5,4 milliards de transistors en tout et fait plus de 4 cm2 de surface. Et surtout, il ne consomme que 70 mW, ce qui permet d'envisager d'empiler ces processeurs en couches, quelque chose d'impossible avec les processeurs CMOS habituels qui consomment beaucoup plus d'énergie.

A titre de comparaison, un processeur Intel Core i7 de dernière génération (Skylake) réalisé en technologie 14 nm consomme entre 15 W et 130 W selon les modèles, pour 1,7 milliards de transistors.

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 85 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.

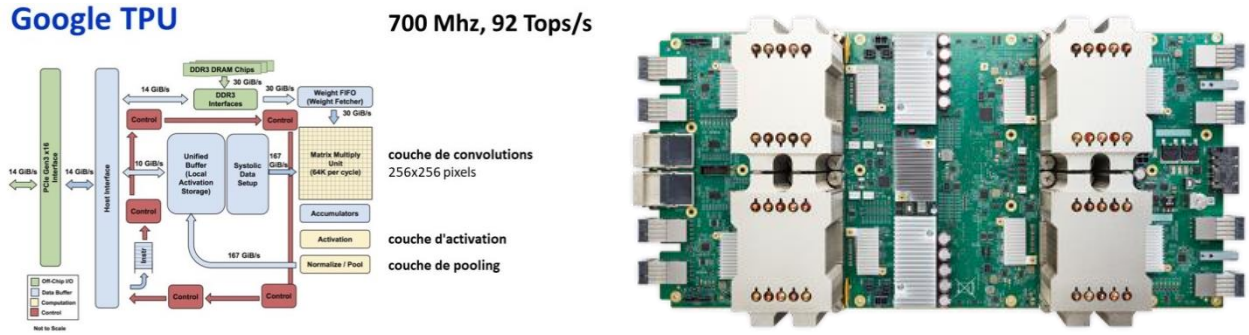
En parallèle, le laboratoire de recherche de Zurich d'IBM planche sur une autre technologie de réseau de neurones[87] à base de GST (Germanium-Antimony-Tellerium) qui est aussi utilisé dans les disques optiques réinscriptibles. Leurs neurones imitent d'encore plus près les neurones biologiques avec une faible consommation, un stockage d'état comme avec les memristors et un peu de fonctionnement aléatoire.





Google a aussi créé ses processeurs neuromorphiques, les 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. Ils en étaient à leur seconde génération à la mi-2017. Ce sont des ASIC performants et consommant peut d'énergie. Leur layout semble surtout adapté à l'exécution de réseaux de convolutions.

[87] Cf https://arstechnica.com/gadgets/2016/08/ibm-phase-change-neurons/.

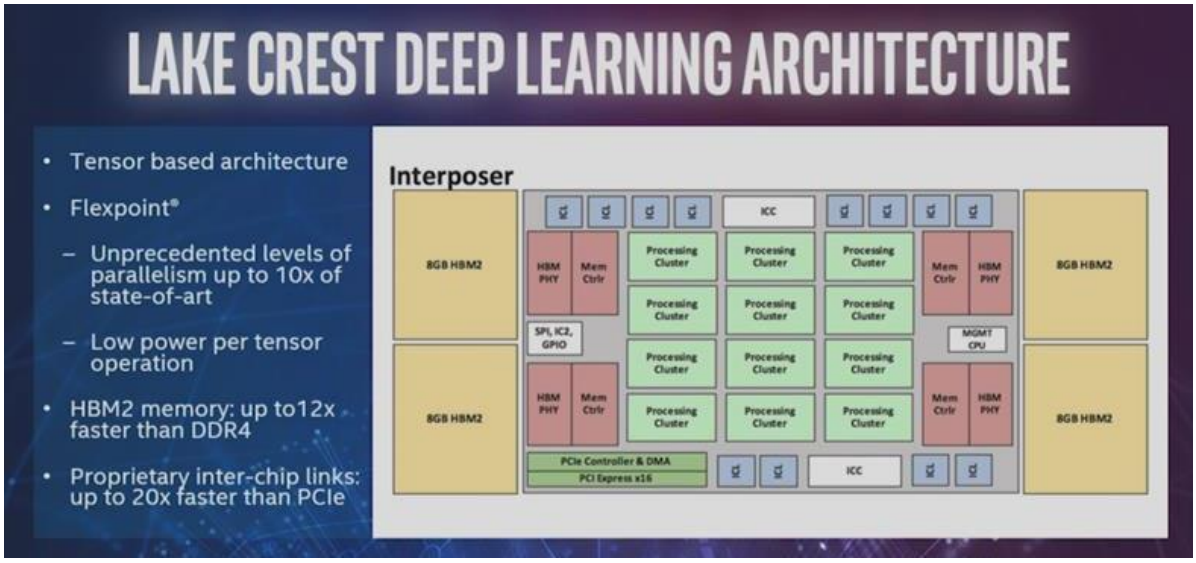


Intel a fait l'acquisition de la startup Nervana ainsi que de Movidius en 2016, lui donnant la capacité de création de processeurs neuromorphiques côté serveur pour le premier et dans l'embarqué pour le second.

Du côté des serveurs, l'offre Nervana est intégrée dans Lake Crest (2016) qui est suivi dans la roadmap Intel par Knights Crest (2017). Ils combinent un processeur Xeon et un coprocesseur neuromorphique développé en technologie FPGA Altera/Intel.

Ce coprocesseur 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.

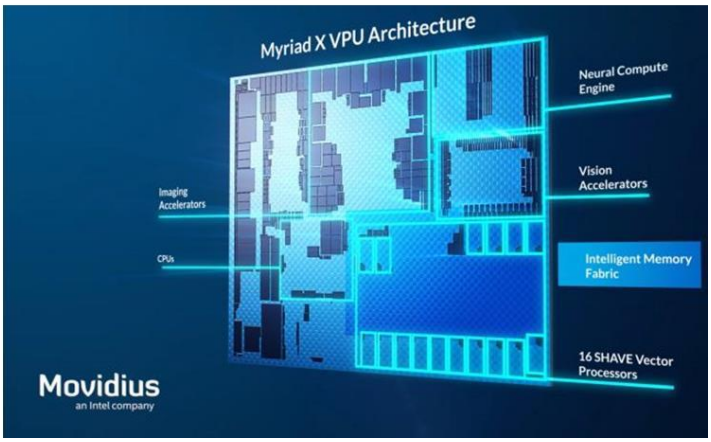
Par contre, Intel n'est pas très bavard sur l'archirecture interne des tenseurs (multiplicateurs de matrices) de ses coprocesseurs ! Intel annoncait la livraison de ses premiers FPGA Nervana Lake Crest à la mi-octobre 2017[88].



Pour l'embarqué, les versions commerciales des chipsets Fathom de Intel Movidius ont été annoncées mi 2017. Ces chipsets exploitant des processeurs vectoriels, dédiés au traitement de l'image. En août 2017, Intel annonçait une nouvelle génération de processeurs Myriad X, remplaçant les Myriad 2.

[88] Cf Intel Shipping Nervana Neural Network Processor First Silicon Before Year End, Anandtech, octobre 2017.

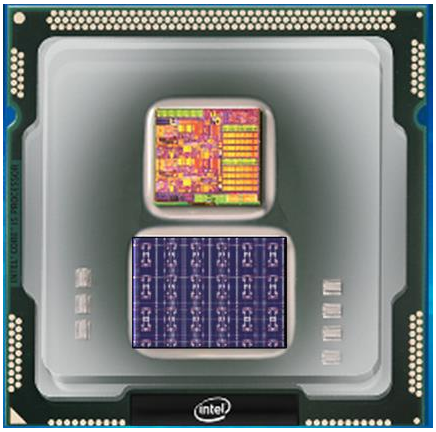
Vision Processing Unit
pour l'embarqué
1 trillion d'ops DNN /s
(x10 vs Myriad 2)
16 processeurs vectoriels
ASIC TSMC 16 nm
encodage vidéo 4K
TDP de 2W



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.

Fin septembre 2017, **Intel** annonçait son processeur neuromorphique Loihi. Il est censé arriver à la mi-2018 sous forme d'un chipset de test pour la recherche.

Il sera fabriqué en technologie 14 nm comme les Core i5/i7 du moment, et comprendra 130 000 neurones impulsionnels, comme dans les chipsets TrueNorth d'IBM. Ces neurones seront reliés entre eux par 130 millions de synapses.



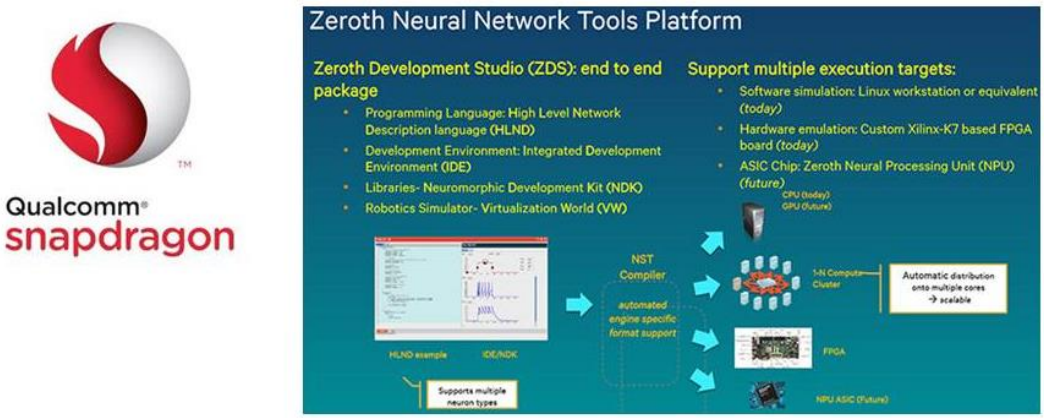
Le marketing d'Intel évoque un processeur imitant le cerveau humain et doué de facultés d'apprentissage, en précisant qu'il supportera des réseaux de neurones récurrents, hiérarchiques et parcimonieux (sparse) et donc en particulier à tout ce qui correspond au traitement du langage et analyse de flux de données temporels divers comme des ECG. 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 entrainés, ici, le processeur est capable d'apprentissage et dans les modes supervisés, non supervisés et par renforcement. Sans qu'Intel ait fourni de détails, on voit dans l'illustration associée ci-dessus que ce processeur comprend en fait 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).

Dans le même temps, Intel fait des efforts pour optimiser les frameworks de deep learning (TensorFlow, Torch, Theano, 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[89] qui équipent les serveurs de data centers, rapprochant leurs performances des meilleurs GPU Nvidia. Bref, Intel aurait du mou sous la pédale dans ses processeurs serveurs !

Qualcomm a conçu son architecture de réseaux de neurones Zeroth intégrée dans sa famille de processeurs Snapdragon mais l'architecture matérielle n'est pas bien claire. Dans la pratique, elle exploite juste un DSP du processeur, le Hexagon, mais ne dispose pas de véritable fonction liée au deep learning. Ce DSP comprend toutefois des unités de manipulation de vecteurs qui optimisent le fonctionnement des réseaux de neurones, comme ceux que l'on trouve dans les processeurs serveur Intel Xeon Phi.



Il en va autrement du chipset Kirin 970 de **HiSilicon**, la filiale de semiconducteurs du chinois Huawei. Présentée à l'IFA, il s'agit d'un chipset mobile gravé en 10 nm par TSMC et comprenant 5,5 milliards de transistors, dont une partie liée au traitement d'applications de deep learning comme la reconnaissance de la parole ou d'images qu'ils appellent une Neural Processing Unit (NPU). Le tout est complété de 8 cœurs CPU et 12 cœurs GPU MALI (design de CPU d'origine ARM). Le NPU peut traiter 1,92 TFlops (en calculs flottants FP16). Le NPU supporte Tensorflow, Tensorflow Lite et Caffe/Caffe2.

[89] Cf 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.

"NPU", ASIC gravé en 10 nm
multiplie des matrices de 3x3
dans Huawei Mate 10 et Mate 10 Pro
2 Tflops/s

dans les iPhone 8 et X
1 Tflops/s

Ce NPU utilise une conception de circuit provenant d'une startup chinoise de Beijing, **Cambricon Technology**. Cette société a été créée en 2016 et vient de lever au mois d'aout la bagatelle de $100M, auprès d'un investisseur public chinois qui ressemble à notre Bpifrance. 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 et notamment pour l'adapter au processus de fabrication du chipset qui est en intégration à 10 nm, fabriqué par TSMC à Taïwan.

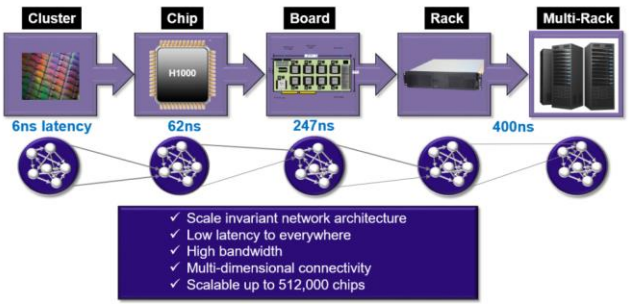
Une semaine après l'annonce du Kirin 970 par Huawei, **Apple** lançait sa nouvelle salve d'iPhones 8 et X. Ceux-ci intègrent aussi une fonction neuromorphique sous la fome d'un coprocesseur dénommé A11 Bionic Neural Engine. Il tourne à 900 MHz mais rien n'a encore filtré sur ses capacités techniques 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. D'autres annonces de ce type suivront très probablement entre fin 2017 et 2018 chez Qualcomm, Samsung et Mediatek.

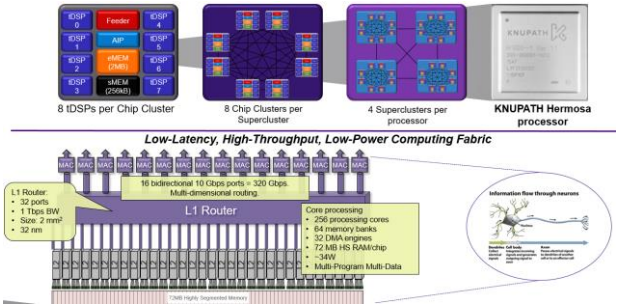
Dans l'embarqué, nous avons aussi des chipsets chez **Brainchip**, **Graphcore** ainsi que chez le français **Chronocam** (chez qui Intel est le plus gros investisseur). Le chipset de Chronocam est en fait un capteur vidéo qui intègre un réseau de neurone permettant l'interprétation immédiate des images. Il existe d'autres projets d'ordinateurs synaptiques à base de réseaux de neurones. On peut notamment citer le projet de **Jeff Hawkins**, le fondateur de Palm et celui de Stanford, qui travaille sur le chipset **Neurocore** intégrant pour l'instant 65536 neurones et fonctionnant à très basse consommation. Les concepteurs de blocs fonctionnels pour processeurs embarqués qui les commercialisent sous forme de propriété intellectuelle s'y sont aussi mis, comme **Tensilica** avec ses Vision C5 surtout destinés aux systèmes de vision artificielle comme les caméras de surveillance.

Créé dans le cadre du projet européen Human Brain Project, le projet **SpiNNaker** de Steve Furber (Université de Manchester, UK) vise à créer un chipset d'un milliard de neurones. Il s'appuie cependant sur une architecture matérielle classique, avec 18 cœurs 32 bits ARM par chip. On est plus dans l'architecture massivement parallèle avec des milliers de processeurs de ce type que dans les processeurs véritablement synaptiques.

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.





La France est aussi présente dans ce créneau.

Le **CEA LETI** planche sur des chipsets neuromorphiques produits en technologie FDSOI par **STMicroelectronics**. Ils sont dédiés aux systèmes embarqués et à la reconnaissance d'images.

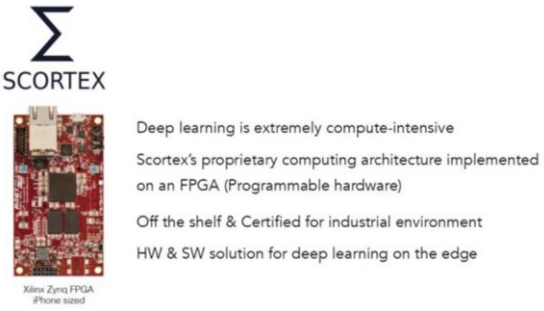
La startup française **Scortex** qui conçoit des FPGA pour des applications de traitement de l'image. Le créateur d'Aldebaran, Bruno Maisonnier, est aussi en phase de lancement d'**Anotherbrain,** une startup de processeur neuromorphique encore stealth qui exploiterait des travaux du Collège de France et de l'Université Pierre et Marie Curie.

CEA N2-D2 DNN : 1.8 TOPS/W 28nm FDSOI

Target	Frequency	Performance	Energy efficiency (SoC only)
Raspberry PI 2 B	900 MHz	480 images/s	380 images/W
Odroid Xu3	2000 MHz	870 images/s	350 images/W
Nvidia K1	**850 MHz**	**3550 images/s**	**600 images /W**
PNeuro (FPGA)	100 MHz	5000 images/s	2000 images/W
PNeuro (FDSOI)	**1000 MHz**	**50000 images/s**	**250000 images/W**

X10

X125



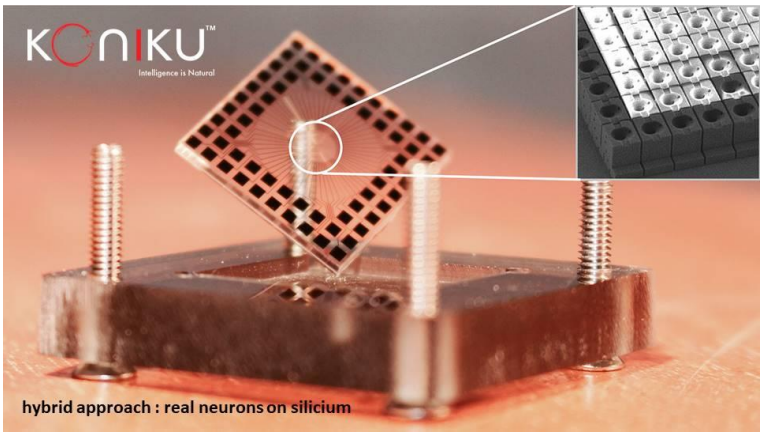
Une équipe de chercheurs associant le **CNRS** et **Thales** située à Palaiseau travaille aussi sur une technologie avancée de processeurs neuromorphiques, en collaboration avec des laboratoires de recherche japonais et américains[90]. Ils utilisent des oscillateurs qui permettent de se rapprocher encore plus du mode de fonctionnement des neurones biologiques en facilitant la propagation temporelle des valeurs entre les neurones d'un système. Cela a l'air de servir surtout aux réseaux de neurones récurrents qui font de la reconnaissance de la parole.

Le **CogniMem** 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

[90] Cf Neuromorphic computing with nanoscale spintronic oscillators, janvier 2017.

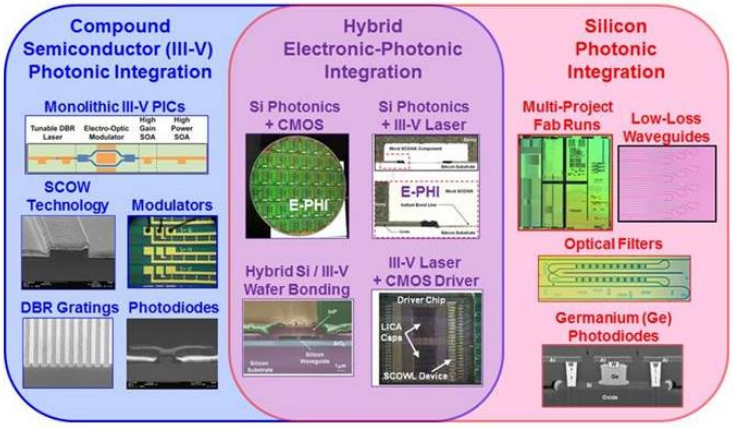
intégrée dans les processeurs d'objets connectés Curie d'Intel (avec 128 neurones, mais abandonnés par ce dernier en juillet 2017). L'ensemble sert principalement aux applications de vision artificielle dans les systèmes embarqués.

Enfin, terminons ce panorama bigaré avec **Koniku** (2014, $1,65M), une startup qui développe des neurones hybrides en silicium et biologiques. La société californienne se positionne aussi sur la reconnaissance d'images. Elle communique peu et il est difficile d'évaluer l'intérêt précis de sa technologie. Mais ça en jette !



Photonique

C'est la photonique qui exploite des composants à base des matériaux dits "III-V"[91]. 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.

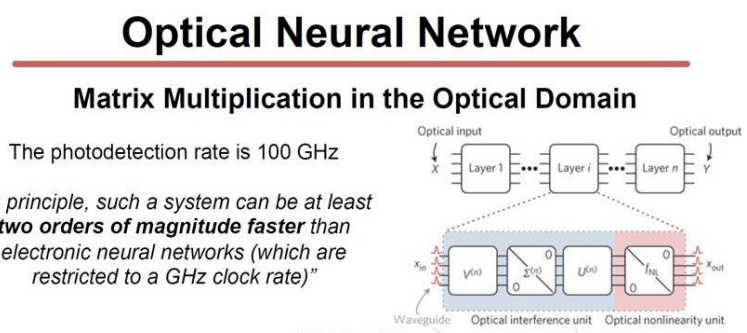


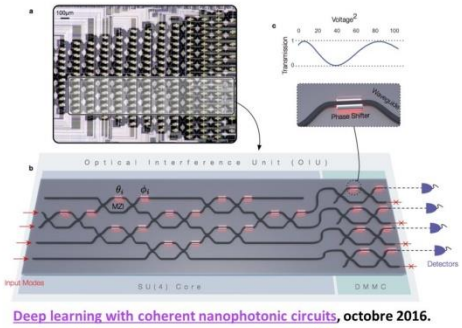
La startup française **Lighton.io** (2016) planche sur la création d'un coprocesseur optique capable de réaliser très rapidement des calculs sur de gros volumes de données et de combinatoires. Le système s'appuie sur la génération de jeux de données aléatoires permettant de tester simultanément plusieurs hypothèses de calcul, à des fins d'optimisation. Les applications visées sont en premier lieu la génomique et l'Internet des objets.

[91] Un sujet que j'avais exploré dans **Comment Alcatel-Lucent augmente les débits d'Internet** en 2013.

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.





Les premiers chipsets expérimentaux photoniques à réseau de neurones ont en tout cas récemment fait leur apparition en laboratoire. Avec à la clé un potentiel de multiplication de la performance par 50[92] !

Ordinateurs moléculaires

Ils permettraient de descendre le niveau d'intégration au-dessous du nanomètre en faisant réaliser les calculs par des molécules organiques de la taille de l'ADN. Cela reste aussi un **animal de laboratoire** pour l'instant ! Mais un animal très prometteur, surtout si l'architecture correspondante pouvait fonctionner de manière tridimensionnelle et plus rapidement que notre cerveau. Reste aussi à comprendre quelle est la vitesse de commutation de ces composants organiques et comment ils sont alimentés en énergie.

Toutes ces innovations technologiques devront surtout se diffuser à un cout raisonnable. En effet, si on extrapole la structure de cout actuelle des superordinateurs, il se pourrait qu'un supercalculateur doté de la puissance du cerveau à une échéance pluridécennale soit d'un cout supérieur au PIB de l'Allemagne[93]. Ca calme ! La puissance brute est une chose, son rapport qualité/prix en est une autre !

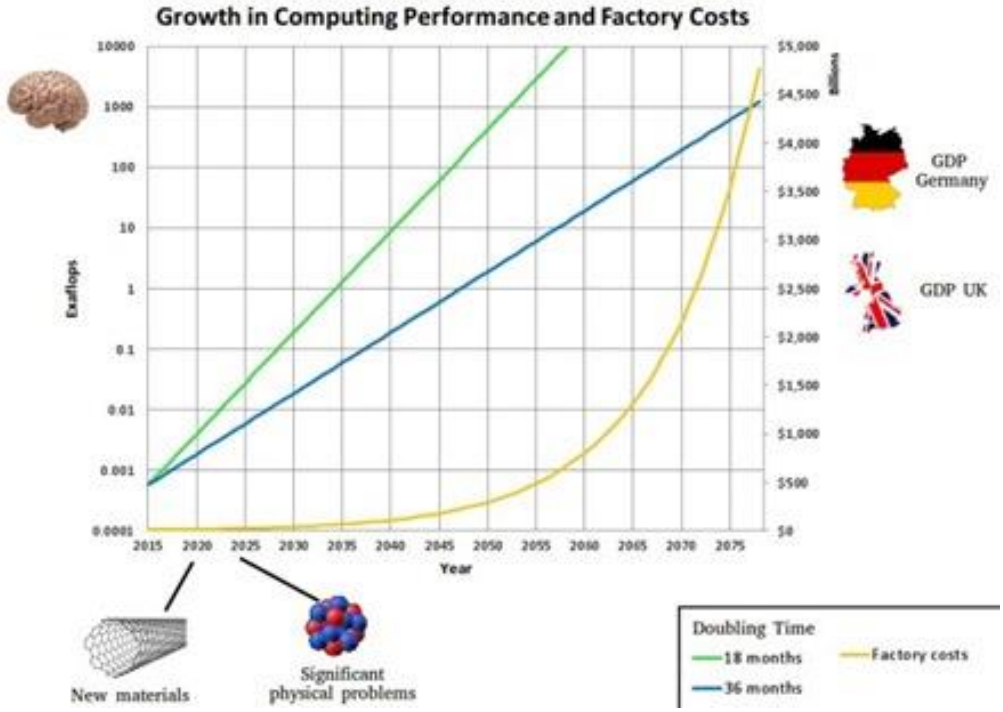
La notion d'IA intégrative pourrait aussi voir le jour dans les architectures matérielles. Comme le cerveau qui comprend diverses parties spécialisées, un ordinateur doué d'IA évoluée intégrera peut-être des architectures hybrides avec processeurs au graphène, optiques et quantiques en compléments d'une logique de base avec de bons vieux CPU en produits en technologie CMOS traditionnelle.

---

[92] Cf Deep learning with coherent nanophotonic circuits, octobre 2016. Voir aussi 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.

[93] Source : Why we need Exascale and why we won't get there by 2020, 2014.

Ceci est d'autant plus plausible que certaines techniques sont insuffisantes pour créer un ordinateur générique, notamment les ordinateurs quantiques qui ne sauraient gérer qu'une certaine classe de problèmes, mais pas comprimer ou décomprimer une vidéo par exemple, ou faire tourner une base de données NoSQL.



Ordinateurs quantiques

Imaginés par le physicien Richard Feynman en 1982, les ordinateurs quantiques sont à même de résoudre certaines classes de problèmes complexes d'optimisation où plusieurs combinatoires peuvent être testées simultanément. Les algorithmes peuvent être résolus de manière polynomiale et non exponentielle. Cela veut dire qu'au gré de l'augmentation de leur complexité, le temps de calcul augmente de manière linéaire avec cette complexité et pas de manière exponentielle. Donc… c'est beaucoup plus rapide !

Le principe des qubits qui sous-tendent les ordinateurs quantiques est décrit dans **Quantum computation, quantum theory and AI** de Mingsheng Ying, qui date de 2009.

Pour faire simple, l'intérêt d'un qubit est de pouvoir avoir plusieurs états probabilistes en même temps, avec ce que l'on appelle la superposition des états. En combinant plusieurs qubits, cela permet de trouver très rapidement un point d'équilibre pour résoudre des équations d'optimisation complexes comprenant de nombreuses variables. Un système à base de n qubits est ainsi capable de représenter simultanément 2^n états !

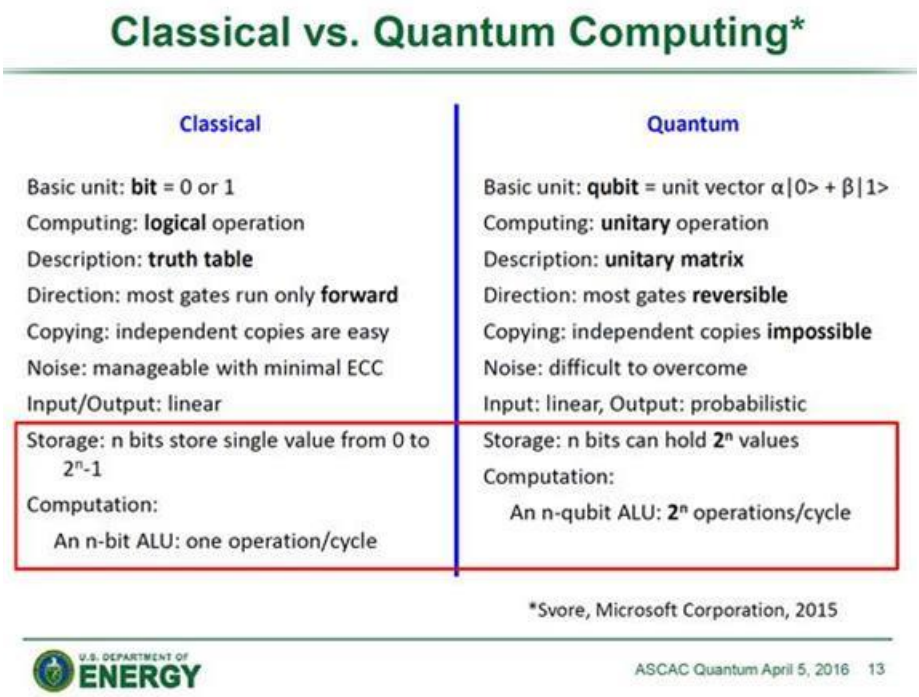
Mais les qubits sont complexes à mettre en œuvre : quelle que soit la technologie employée, ils doivent être refroidis à quelques dizaines de millikelvins au-dessus du zéro absolu avec des systèmes à base d'hélium liquide. Il est surtout difficile d'initialiser leur état et de le lire.

L'un des premiers algorithmes 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 son exponentielle comme avec les calculateurs traditionnels. Il permet de casser les clés publiques utilisées en cryptographie avec l'algorithme RSA. Ce qui le remet sérieusement en question ! 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.

Dans **Quantum POMPDs**, Jennifer Barry, Daniel Barry et Scott Aaronson, du MIT, évoquent en 2014 comment les ordinateurs quantiques permettent de résoudre des problèmes avec des processus de décision markovien partiellement observables. Il s'agit de méthodes permettant d'identifier des états optimaux d'un système pour lequel on ne dispose que d'informations partielles sur son état.

Quant à **Quantum Speedup for Active Learning Agents**, publié en 2014, un groupe de scientifiques espagnols et autrichiens y expliquent comment les ordinateurs quantiques pourraient servir à créer des agents intelligents dotés de facultés d'auto-apprentissage rapide. Cela serait un chemin vers le développement de systèmes d'IA créatifs.

En 2014, des chercheurs chinois de l'Université de Sciences et Technologies de Hefei ont été parmi les premiers à **expérimenter** des ordinateurs quantiques pour mettre en jeu des réseaux de neurones artificiels, pour la reconnaissance d'écriture manuscrite. Leur ordinateur quantique utilise un composé organique liquide associant carbone et fluor. De nombreuses autres publications font état depuis de la possibilité d'utiliser des ordinateurs quantiques pour entrainer rapidement des réseaux de neurones[94].



[94] Comme Application of Quantum Annealing to Trainingof Deep Neural Networks, de Adachi et Henderson, 2015.

Les nombreuses différences entre un ordinateur classique et un ordinateur quantique sont résumées *ci-dessus*. Avec quelques nuances de taille comme l'impossibilité de copier l'état d'un qubit sur un autre qubit.

Il existe de nombreuses catégories de processeurs quantiques qui se définissent par leur technologie de qubits. Les principales sont 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, même si leur efficacité est contestée. Pour les puristes, ce ne sont pas véritablement des ordinateurs quantiques mais plutôt des simulateurs d'ordinateurs quantiques.
- **Boucles supraconductrices**, ou superconducting loops, chez **IBM** et **Google** ainsi qu'au CEA en France.
- **Qubits topologiques**, chez **Microsoft** avec les fermions de Majorana dont l'existence vient tout juste d'être prouvée en laboratoire, et dans les **Bell Labs** de Nokia.
- **Quantum dots sur silicium**, chez **Intel**, qui vient d'en faire une annonce officielle en octobre 2017 ainsi qu'au **CEA**.
- **Ions piégés**, comme chez la startup **ionQ** (2016, $20M) qui s'appuie sur des travaux de l'Université du Maryland et de l'Université Duke en Caroline du Nord. Elle ambitionne de sortir son ordinateur quantique en 2018, ce qui est probablement plus qu'optimiste.

Ces techniques que nous n'aurons pas le temps ni le courage de décrire ici en détails ont chacune leurs avantages et inconvénients. Certains types de qubits sont notamment plus difficiles à stabiliser que d'autres.

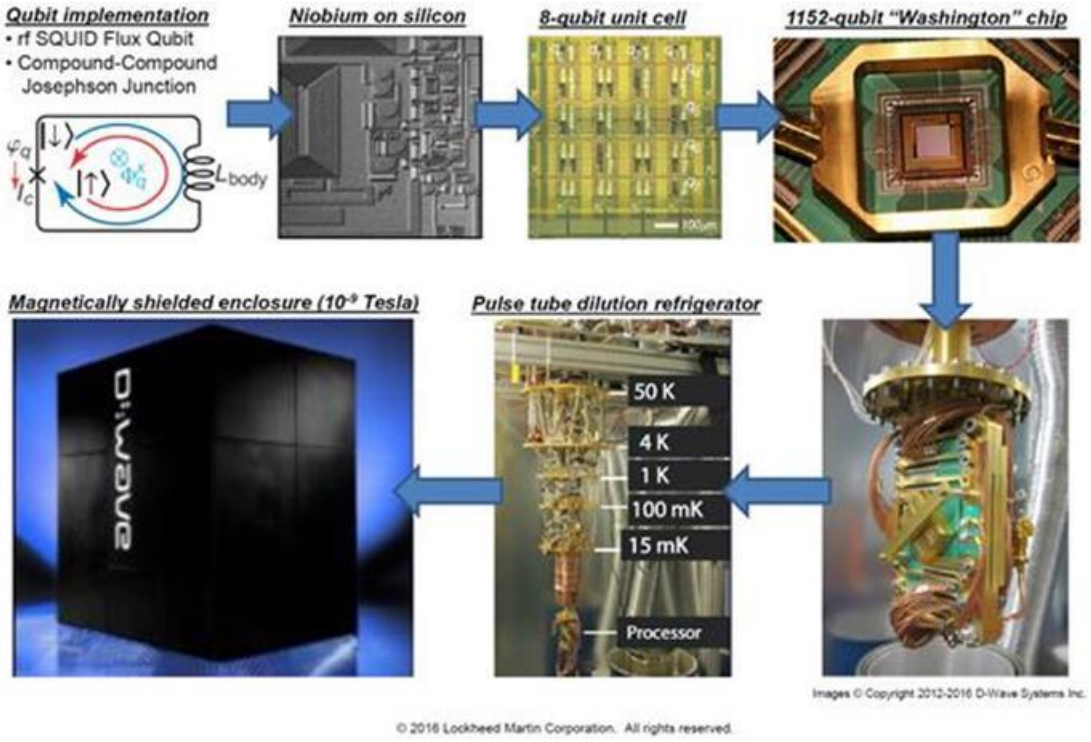
Voyons tout cela dans l'ordre…

D-Wave est la société la plus avancée dans le domaine du quantique avec ses ordinateurs dont les capacités évoluent régulièrement depuis une demi-douzaine d'années. Ils sont commercialisés à petite échelle. Leur dernier D-Wave 2 a une capacité de 2031 qubits, un record en la matière. Leurs qubits sont fabriqués à base de niobium sur silicium, utilisant l'effet Josephson. La partie quantique est isolée magnétiquement de l'extérieur, avec un champ interne d'un nano-Tesla.

Elle réalise une simulation toutes les 25 µs, dans des batches de 10 000 opérations identiques durant 4 secondes. La moyenne des résultats de ces opérations est alors évaluée. L'ordinateur consomme 15KW.

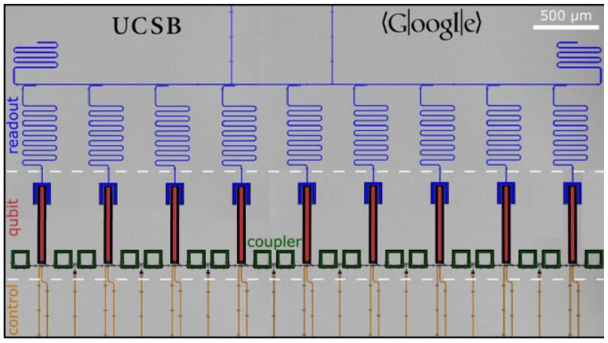
Les ordinateurs de D-Wave sont notamment utilisés par les équipes de la NASA dans leur QuAIL, le **Quantum Artificial Intelligence Laboratory**, un laboratoire de recherche lancé en partenariat avec Google Research. Il utilise un D-Wave Two comme outil d'expérimentation. Leurs **publications scientifiques** sont abondantes mais pas faciles d'abord comme les autres ! Ce centre de la NASA est situé au Ames Research

Center, là-même où se trouve la Singularity University et à quelques kilomètres du siège de Google à Mountain View.



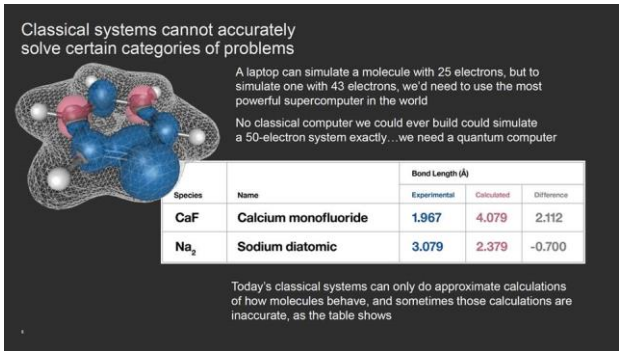
Google annonçait fin 2015 avoir réussi à réaliser des calculs quantiques 100 millions de fois plus rapidement qu'avec des ordinateurs classiques sur ce DWave-Two. Ces tests sont mal documentés au niveau des entrées, des sorties et des algorithmes testés.

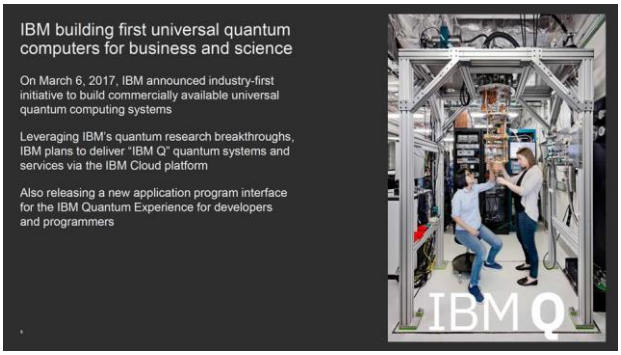
Qui plus est, la comparaison faite par Google avec les calculs sur ordinateurs traditionnels s'appliquait à algorithme identique alors que les algorithmes utilisés dans l'ordinateur quantique n'étaient pas optimisés pour ordinateurs traditionnels. Le sujet est polémique, comme le rapportent **La Tribune** ou **Science et Avenir**. Est-ce une querelle entre anciens et modernes ? Pas vraiment car ceux qui doutent des performances du D-Wave travaillent aussi sur les ordinateurs quantiques.



Google travaille aussi sur la création de ses propres processeurs quantiques à base de boucles superconductrices. Avec l'Université de San Barbara en Californie, il annonçait début octobre 2017 avoir pu créer des composants de 9 qubits stables, générant un taux d'erreur stable et ouvrant la voie à la création d'ordinateurs quantiques opérationnels en 2018[95] (*ci-dessus à droite*). Ils prévoient de créer un chipset équivaleur de 50 qubits.

Début mai 2016, **IBM** annonçait mettre à disposition son ordinateur quantique expérimental cryogénique de 5 Qubits en ligne dans son offre de cloud (*ci-dessus*). On ne sait pas trop quel type de recherche pourra être menée avec ce genre d'ordinateur ni quelles APIs sont utilisées. En 2017, ils en sont à créer des systèmes à base de 17 Qubits. Ils proviennent du laboratoire de Zurich d'IBM.





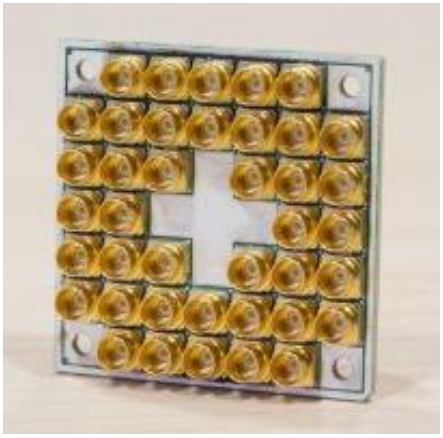
De son côté, **Microsoft** avance aussi très bien sur l'informatique quantique avec une technologie à base de fermion de Majorana[96] qui présente l'intérêt de mieux stabiliser les qubits et qui pourrait monter plus rapidement en puissance.

Ces travaux de recherche sont menés à StationQ, un laboratoire de Microsoft situé à Santa Barbara en Californie et dirigé par Michael Friedman, un mathématicien médaille Fields. L'éditeur a de plus déjà développé les briques logicielles pour créer des applications autour de ses ordinateurs quantiques, dont le langage de programation LIQUiD.

En octobre 2017, **Intel** se lançait officiellement dans la course de l'ordinateur quantique en annonçant un processeur quantique à 17 qubits (vidéo). Il est conçu par des équipes d'Intel siuées en Oregon et en Arizona, en partenariat avec l'institut de recherche en informatique quantique **QuTech** de l'Université de Delft aux Pays-Bas, qui va se consacrer au programme de test et d'évaluation du chipset. Ces qubits sont de la catégorie des supraconducteurs. Ils sont bien isolés les uns des autres et leur connectique est visible sur la photo ci-dessous. Le chipset fait la taille d'une pièce de 50c et est fabriqué sur des wafers de 300 mm.

[95] Voir A blueprint for demonstrating quantum supremacy with superconducting qubits, septembre 2017. Deux tiers de la vingtaine de chercheurs qui ont signé ce papier travaillent chez Google.

[96] Cette particule n'a pas de masse ni d'énergie et est sa propre antiparticule. Son existence a été théorisée en 1937 et démontrée en 2016.



Le **CEA** de Saclay planche depuis longtemps sur la création de circuits quantiques. Ils ont développé en 2009 un **dispositif de lecture d'état quantique** non destructif de qubits après avoir créé l'un des premiers qubits en 2002.

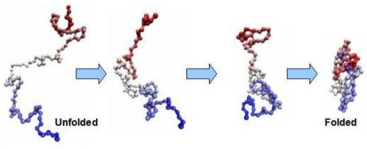
Et le CEA-LETI de Grenoble a de son côté récemment réalisé des **qubits sur composants CMOS** grâce à la technologie SOI d'isolation des transistors sur le substrat silicium des composants.

Enfin, le groupe français **ATOS**, déjà positionné dans le marché des supercalculateurs depuis son rachat de Bull, travaille avec le CEA pour créer un ordinateur quantique **à l'horizon 2030** dans le cadre du programme ATOS Quantum qui et à vocation européenne. En juillet, ATOS annonçait le lancement d'un simulateur d'ordinateur quantique de 30 à 40 qubits, le Atos Quantum Learning Machine, réalisé à base de chipsets Intel CMOS traditionnels, probablement des processeurs de serveurs de type Xeon[97], dont le nombre va de 2 à 16, avec une mémoire allant de 1 à 24 To. Ce calculateur est programmable avec le langage spécifique aQasm (Atos Quantum Assembly Language). L'idée est de se faire la main sur les techniques de programmation d'ordinateurs quantiques avant que ceux-ci ne voient le jour.

Mais la France n'est pas seule en Europe sur l'informatique quantique. Les pays qui semblent le plus en pointe sont la Suisse et les Pays-Bas, dont les laboratoires de recherche travaillent respectivement avec IBM et Microsoft.

exemples d'applications du quantique

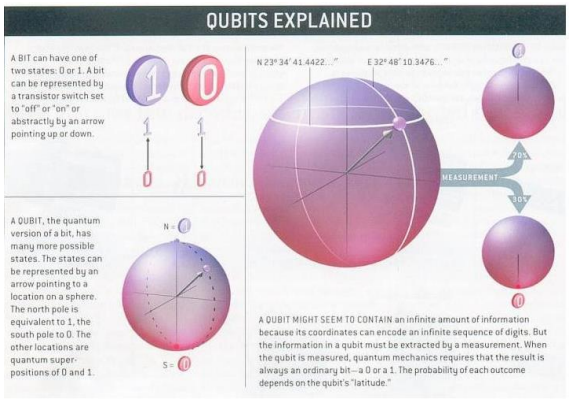
optimisations complexes
simulation de physique des matériaux
simulation de la photosynthèse
cryptographie
simulations statistiques (Monte Carlo)
recherches complexes
génômique combinatoire
criblage en biologie moléculaire
simuler le repliement de protéines

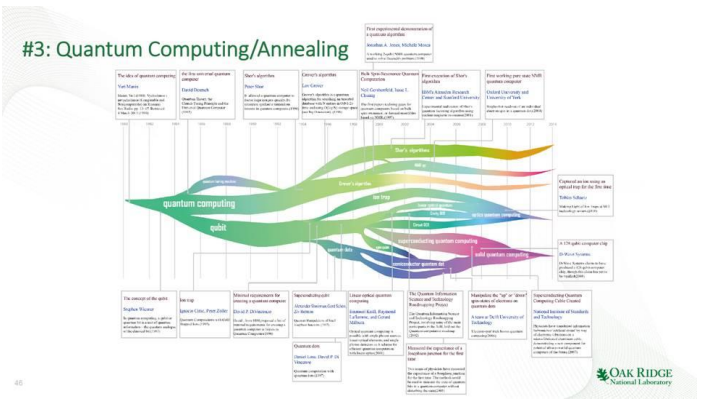


[97] Source : Atos lance aujourd'hui le simulateur quantique le plus performant au monde, juillet 2017.

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. Ils pourront notamment servir à optimiser rapidement certains réseaux de neurones, à faire des simulations physiques dans les matériaux ou dans la biologie moléculaire.

Dans son étude **Quantum Computing Market Forecast 2017-2022**, le cabinet Market Research Media prévoit que le marché des ordinateurs quantiques fera $5B d'ici 2020, en intégrant toute la chaine de valeur matérielle et logicielle. Le premier marché serait celui de la cryptographie. Avant de parler de marché, il faudrait que cela marche ! Et nous n'y sommes pas encore.





Chaque chose en son temps : la recherche, l'expérimentation puis l'industrialisation. Nous n'en sommes qu'aux deux premières étapes pour l'instant.

Mémoire

Les GPU et les processeurs neuromorphiques sont d'autant plus performants dans les phases d'entrainement 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.

Citons les principaux niveaux de mémoire qui équipent aujourd'hui les serveurs.

Mémoire cache

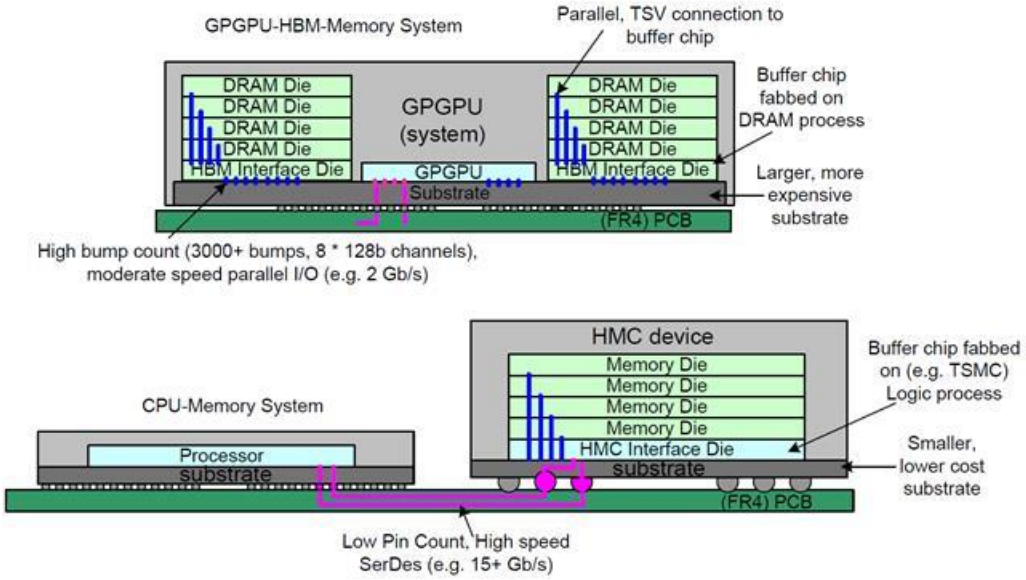
Au sein des processeurs se trouve de la **mémoire cache** qui est utilisée directement par les unités de traitement. Sa vitesse d'accès est ce qui se fait de plus rapide, et dépasse les To/s (téra-octets par secondes).

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.

HBM2 et HMC

Autour des GPU 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).



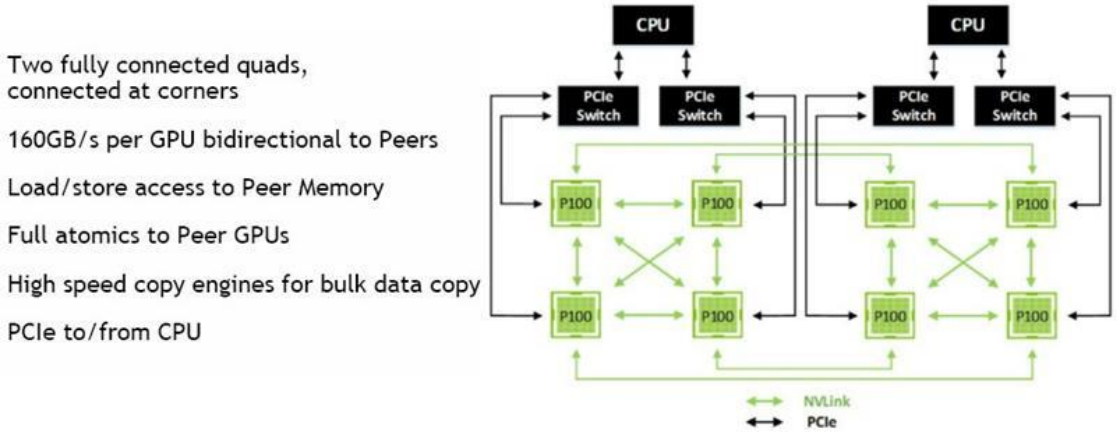
Le premier standard est promu par AMD et le Coréen SK Hynix et le second par l'Américain Micron avec le support de Samsung. Cette mémoire qui atteint aujourd'hui 16 Go est située dans des circuits intégrés empilés pas paquets de 4 ou 8 et reliés entre eux et avec le GPU ou le CPU par des micro-conducteurs métalliques.

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 que le CPU[98]. Ces mémoires permettent d'atteindre des débits très élevés allant jusqu'à 900 Go/s dans le Nvidia GV100 lancé en 2017. HBM2 est utilisé dans les GPU Nvidia GV100 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. 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 !

[98] Source du schéma qui suit : A Talk on Memory Buffers, Inphi.



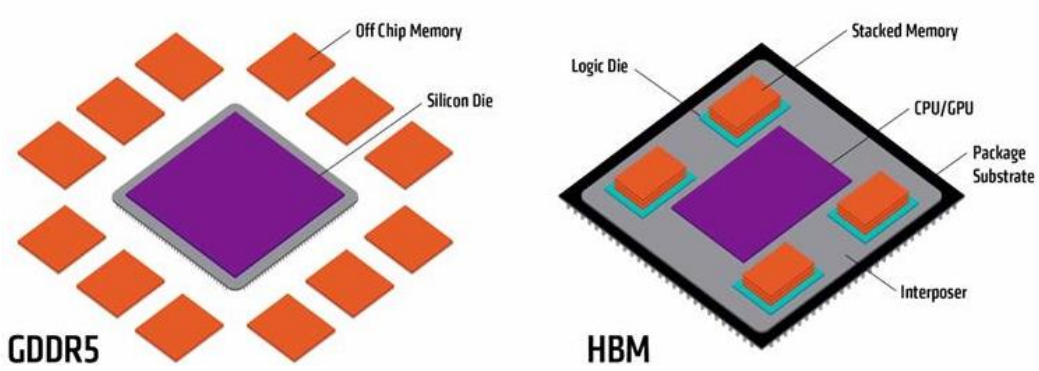
Infiniband

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.

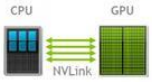
Les composants supportant Infiniband sont commercialisés par l'israélien **Mellanox Technologies** (1999, $89M) et par Intel. Infiniband est notamment complété par le standard RoCE (RDMA over Converged Ethernet) qui permet de permettre l'accès à la mémoire d'un serveur par un autre serveur. Infiniband 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).

GDDR5

La mémoire **GDDR5** 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.



Elle est remplacée par de la mémoire HMC ou HBM2 depuis quelques temps dans les GPU haut de gamme.

		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	
GDDR5 mémoire externe GPU		48 Go/s	2 Go – 12 Go
NVLink comm inter-GPU/CPU		160 Go/s	
HBM2 / HMC mémoire externe GPU		900 Go/s	16 Go
GPU cache & registres mémoire interne GPU		> 16 To / s	6 Mo (L1)

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 maitrise de la fabrication de circuits intégrés de grande taille et avec des dizaines de milliards de transistors.

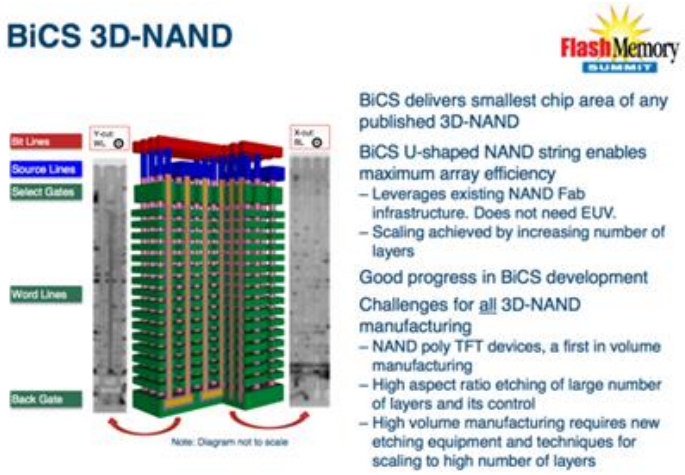
Citons pour terminer sur la mémoire un cas original avec la startup fabless **Upmem** (2015, $3,5M), basée à Grenoble, la capitale française des nanotechnologies, qui intègre des unités de traitement DPU (DRAM Processing Units) directement à l'intérieur de mémoires DRAM, permettant une accélération de 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 !

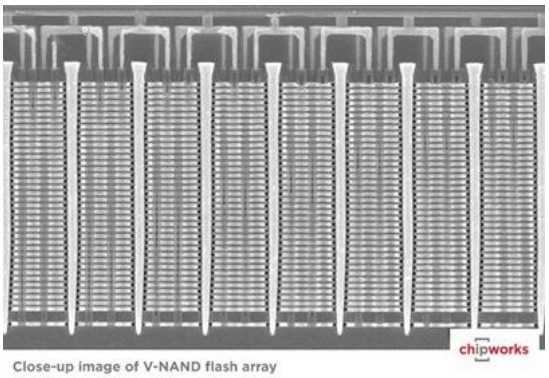
Stockage

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 1 To avec une vitesse d'accès de 3 Go/s. Il existe des SSD de bien plus grande capacité qui sont destinés aux serveurs de data-centers.

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.





Close-up image of V-NAND flash array

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 connait un double retard à l'allumage : côté disponibilité comme côté performance.

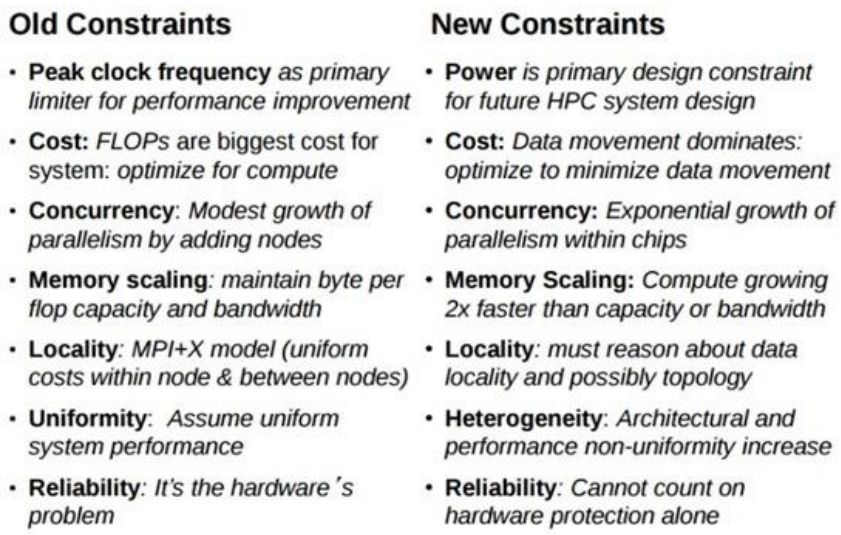
Des disques SSD de 16 To devraient arriver d'ici peu ! Pourquoi cette intégration verticale est-elle possible pour la mémoire et pas pour les processeurs (GPU, CPU) ? C'est lié à la résistance à la montée en température. Dans un processeur, une bonne part des transistors fonctionne en même temps alors 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 100 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 [99]!

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'un des techniques que nous avons examinées juste avant.



Old Constraints	New Constraints
• **Peak clock frequency** *as primary limiter for performance improvement*	• **Power** *is primary design constraint for future HPC system design*
• **Cost:** *FLOPs* are biggest cost for system: *optimize for compute*	• **Cost:** *Data movement dominates: optimize to minimize data movement*
• **Concurrency**: *Modest growth of parallelism by adding nodes*	• **Concurrency:** *Exponential growth of parallelism within chips*
• **Memory scaling**: *maintain byte per flop capacity and bandwidth*	• **Memory Scaling:** *Compute growing 2x faster than capacity or bandwidth*
• **Locality**: *MPI+X model (uniform costs within node & between nodes)*	• **Locality**: *must reason about data locality and possibly topology*
• **Uniformity**: *Assume uniform system performance*	• **Heterogeneity**: *Architectural and performance non-uniformity increase*
• **Reliability**: *It's the hardware's problem*	• **Reliability**: *Cannot count on hardware protection alone*

(source du slide ci-dessus)

Des chercheurs d'université et même chez Microsoft Research cherchent à **stocker l'information dans de l'ADN**. Les premières expériences menées depuis quelques années sont prometteuses[100]. 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

[99] 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 !

[100] 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 !

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.

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 se faire sentir même 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.

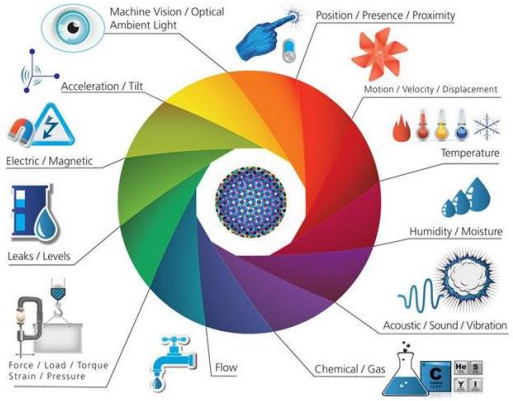
C'est l'une des raisons pour lesquelles les véhicules automatiques sont à terme très prometteurs et plus sécurisés. Ces techniques sont déjà meilleures que les sens humains, surtout en termes de temps de réponse, de vision à 360° et de capacité d'anticipation des mouvements sur la chaussée (piétons, vélos, autres véhicules). A contrario, la finesse de la vue humaine n'est pas encore égalée par la vision artifi-

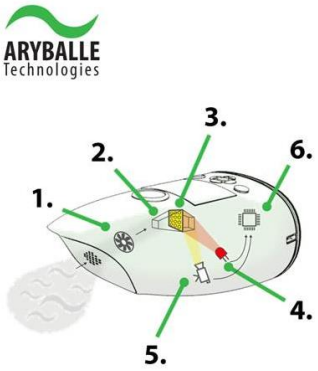
cielle de fait de ses contraintes actuelles. En effet, les réseaux de neurones convolutionnels utilisent des images sources à basse résolution pour tenir compte des contraintes matérielles actuelles.

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. 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, avec le détecteur de gaz du français **Aryballe** ou encore avec le détecteur de pollution aérienne d'un autre français, **Plume Labs**. 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.





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.

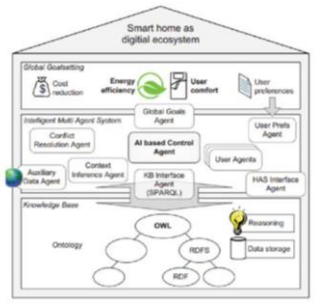
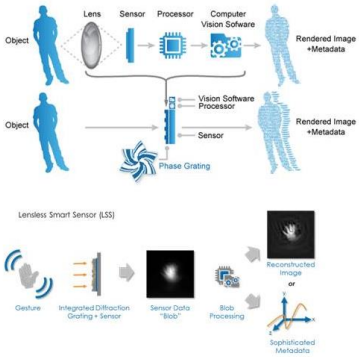


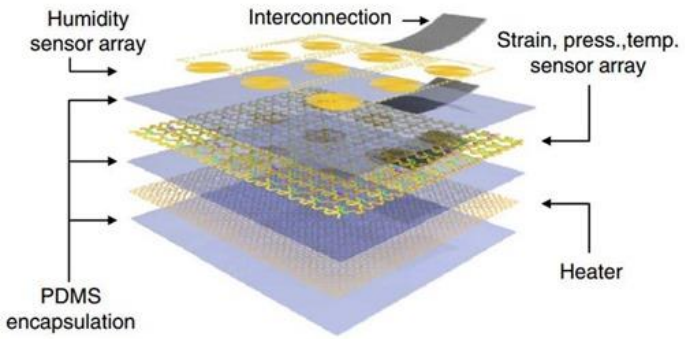
Figure 6.6. The "ThinkHome" smart home project, (Reinisch et al., 2010).



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. 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** installée à Angers depuis 2016 a créé un capteur intégrant un logiciel à base de réseaux de neurones servant à détecter les vibrations anormales.

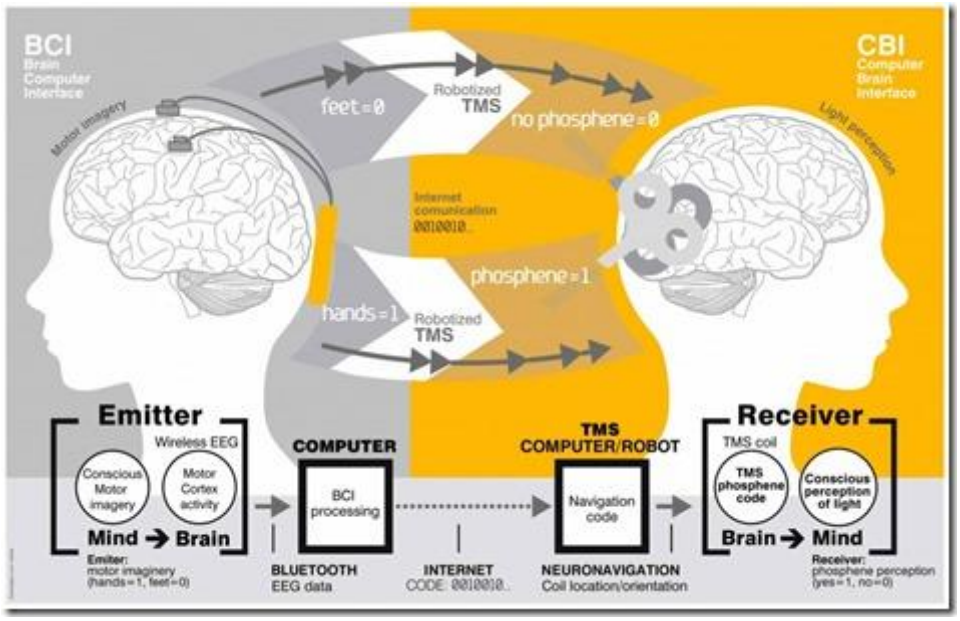
Les capteurs de proximité intégrables à des machines comme les robots progressent même dans leur biomimétisme. Des prototypes de peau artificielle sensible existent déjà en laboratoire, comme en Corée du Sud (*ci-dessous*, **source dans Nature**). L'une des mécaniques humaines les plus difficiles à reproduire 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 muscu-

laires. 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.

Certains se lancent même 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.



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 ou de **OpnWtr**[101] avec son bonnet utilisant des capteurs photos et des émetteurs infrarouges pour cartographier finelement l'état des neurones et même, à terme, le modifier. **Facebook** essaye aussi de lire dans les pensées pour remplacer les claviers[102] !

[101] 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/.

[102] Nanalyze a identifié 29 startups qui s'attaquent à l'interface cerveau-machine dans 29 Neurotech Companies Interfacing With Your Brain, octobre 2017.

Quant au projet européen **Human Brain Project** piloté par le suisse Henri Markram, il ambitionne carrément de cartographier à bas niveau le cerveau humain pour l'émuler ensuite sur ordinateur. Le projet comprend la participation du laboratoire Neurospin du CEA qui met actuellement en place le système d'imagerie par résonnance magnétique nucléaire le plus puissant du monde.

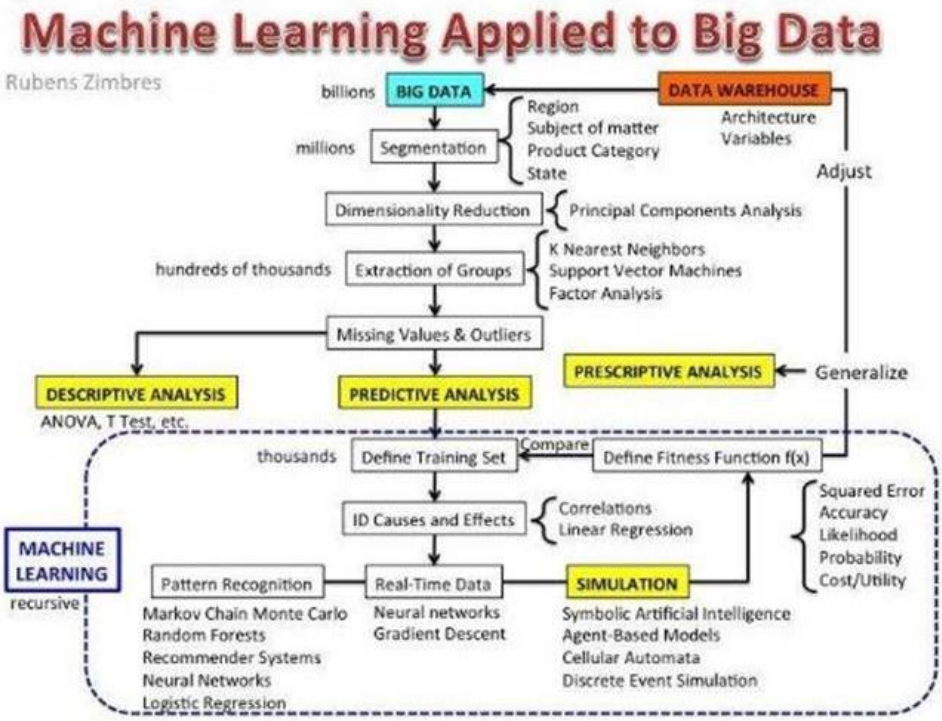
Il faut évidement prendre des pincettes avec tous ces effets d'annonces. Ainsi, dans **Mashable**, une certaine Marine Benoit affirmait un peu rapidement en mars 2016 qu'une équipe avait mis au point "*un stimulateur capable d'alimenter directement le cerveau humain en informations*". A ceci près que l'étude en question, **Frontiers in Human Neuroscience** ne faisait état que d'un système qui modulait la capacité d'acquisition par stimulation ! 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. Mais ce n'est peut-être qu'un début !

Big data

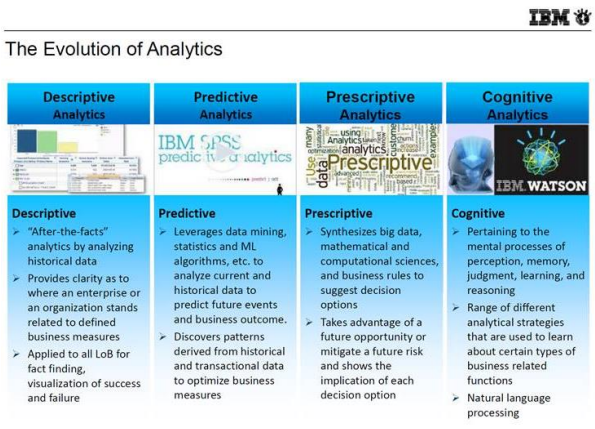
Le domaine du big data est étroitement lié à l'intelligence artificielle. C'est même d'une certaine manière le sang qui l'alimente. Nombre d'applications de machine learning et de deep learning exploitent de gros volumes de données internes aux entreprises. Plus l'entreprise détient de points de contacts avec des infrastructures ou des clients, plus volumiques sont les données captées exploitables.

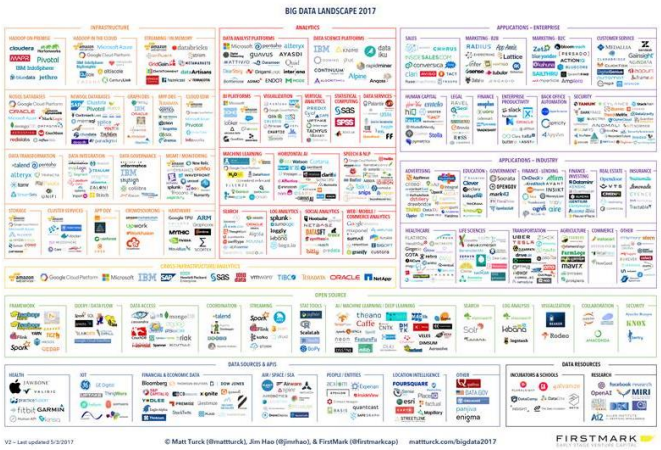
Les infrastructures de big data sont donc clés pour alimenter les applications de l'IA. Les données consolidées étaient peut-être faiblement exploitées jusqu'à présent et le seront mieux grâce à la puissance des outils de l'IA, surtout le machine learning dans un premier temps. Celui-ci est d'ailleurs souvent présenté comme l'aboutissement des outils de *data analytics* et de *data intelligence*.

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 (cf la cartographie *ci-dessous*).





Cloud

Les applications de l'IA font aussi appel aux ressources du cloud, en particulier dans les phases d'entrainement et surtout pour les startups qui ne peuvent pas disposer de leur propre data center.

Les grandes entreprises auront à gérer un équilbre 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'entrainement de modèles de machine learning et deep learning nécessitant d'aligner parfois des milliers de serveurs. Une fois les modèles entrainés, 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 entrainement !

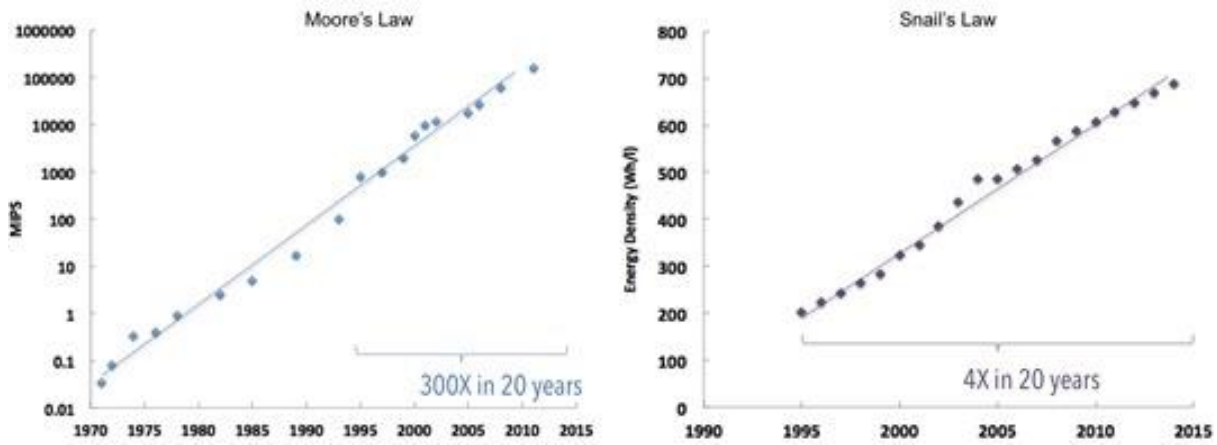
Energie

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) 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.

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[103].

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.

[103] Cf http://blog.erios.org/index.php?post/2013/12/07/Stockage-de-l-%C3%A9lectricit%C3%A9%3A-les-batteries-du-futur-face-au-tout-p%C3%A9trole.

Applications génériques de l'IA

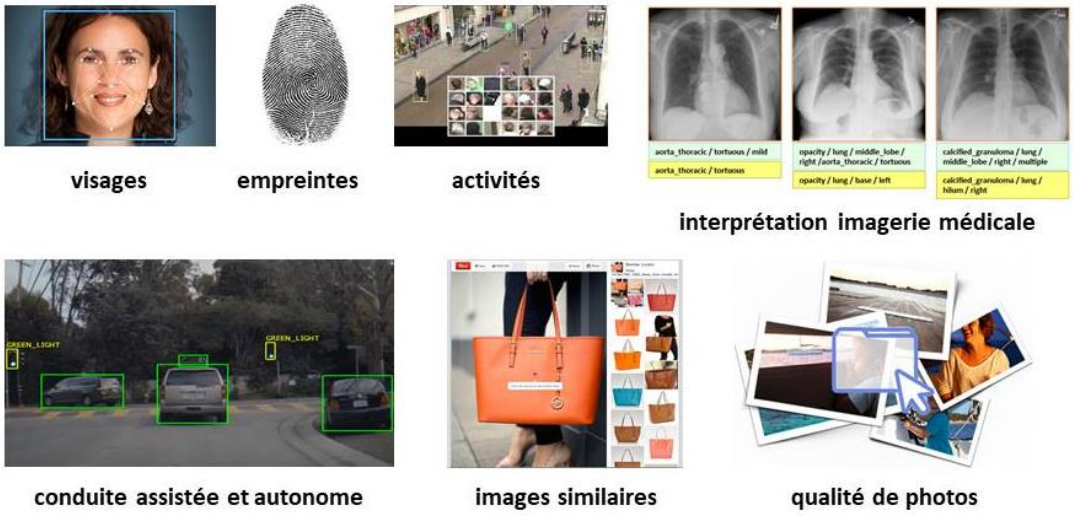
Cette partie est dédiée aux applications génériques de l'IA avec deux types :

- La vision, le langage et la robotique qui s'appuient sur les briques fondamentales de l'IA vues précédemment.
- Le marketing, les ressources humaines et la cybersécurité qui font appel aux briques technologiques de l'IA ainsi qu'aux trois domaines précédents, selon les besoins.

Elles sont potentiellement mises en œuvre par les entreprises de tous secteurs et de toutes tailles.

Vision

La vision artificielle est l'application la plus courante et diversifiée de l'IA. C'est l'une des principales applications des réseaux neuronaux et du deep learning. L'un des objectifs de la recherche est d'élever au maximum le niveau sémantique de la reconnaissance, pour identifier les personnes et objets sur les images.

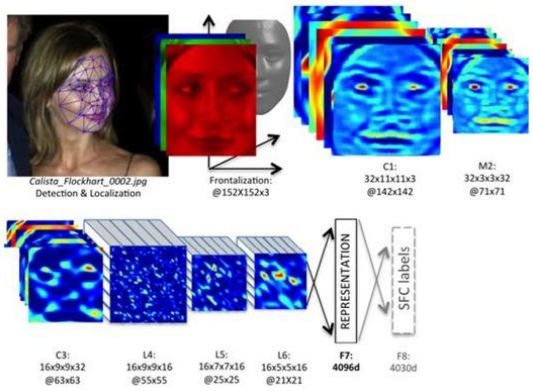


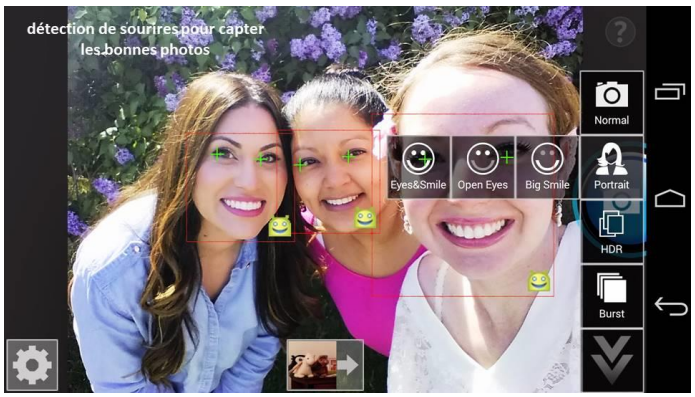
Nous allons passer en revue quelques-uns des principaux usages de la vision artificielle.

Reconnaissance de visages

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).

Dans leur projet FaceNet, Google annonçait en 2015 avoir atteint un taux de réussite de détection de visage de 99,63%[104]. Le tout s'appuyait sur un réseau neuronal à 22 couches.





De son côté, **Facebook** et son projet DeepFace s'appuyait sur la technologie issue d'une start-up israélienne **face.com**. Son taux de réussite était 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 juste en-dessous du taux de reconnaissance humain qui serait évalué à 97,5%.

affectiva®

Microsoft

détection d'émotions

login par reconnaissance de visages
avec Windows Home et Intel RealSense

On trouve de la détection de visages dans plein de solutions du marché comme avec la fonction Faces d'**Apple** iPhoto. Elle provient peut-être de la start-up suédoise **Polar Rose** acquise par Apple en 2010. De manière peu surprenante, Apple a aussi acquis, début 2016, la start-up **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.

Les APIs en cloud proposées par Microsoft Research dans le cadre de son **projet Oxford** apportent des services équivalents aux développeurs d'applications. **Google** fait de même avec ses **Cloud Vision APIs**. 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.

[104] Voir FaceNet: A Unified Embedding for Face Recognition and Clustering. publié en juin 2015.

Il faut ensuite les mettre en œuvre avec du logiciel et du matériel. La différence se situe dans l'implémentation 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 et 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, $642K) et ses API en cloud de tagging automatique d'images en fonction de leur contenu et **Cognitec** qui vise surtout les marchés de la sécurité. Nous avons aussi la startup française **Smart Me Up** (2012, 3m€), 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.

En Chine, **SenseTime** (2014, $410M) 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 nombreuses : suivi de plusieurs visages en temps réel dans des vidéos, 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).

Les solutions de reconnaissance de visage qui évaluent l'âge sont généralement à côté de la plaque. L'habit (du visage) ne faisant pas le moine (qui est quinqua) ! Ici, sur le stand de Smart Me Up au CES 2016 de Las Vegas.

On trouve des solutions de reconnaissance de visage dans les vidéos chez **Kairos** qui savent aussi analyser les émotions et quantifier les foules, chez **KeyLemon** ($1,5m de levés) qui propose une solution en cloud, chez **Matroid** (2016, $3,5M), qui fonctionne sur des flux vidéo ou des photos, chez **Clarifai** ($10m de levés) qui permet notamment de faire de la curation de contenus photo et vidéo et d'ajouter une fonction de recherche d'image par tags ou similaires dans son site, ou chez le japonais **NEC**. Il faut aussi citer **OpenCV**, une solution open source de détection de vi-

sages. Voir **cette liste** de solutions pour développeurs de détection de visages dans les vidéos.

La reconnaissance de visages sert évidemment aussi aux applications de vidéo-surveillance, comme celle de **Camio** (2013) qui fournit une solution en cloud d'exploitation de vidéos de caméras de surveillance.

En complément de la reconnaissance d'images, on peut aussi les modifier et les améliorer. C'est ce que propose **Adobe** avec Sensei, une application qui corriger les perspectives et divers paramètres de selfies (vidéo).

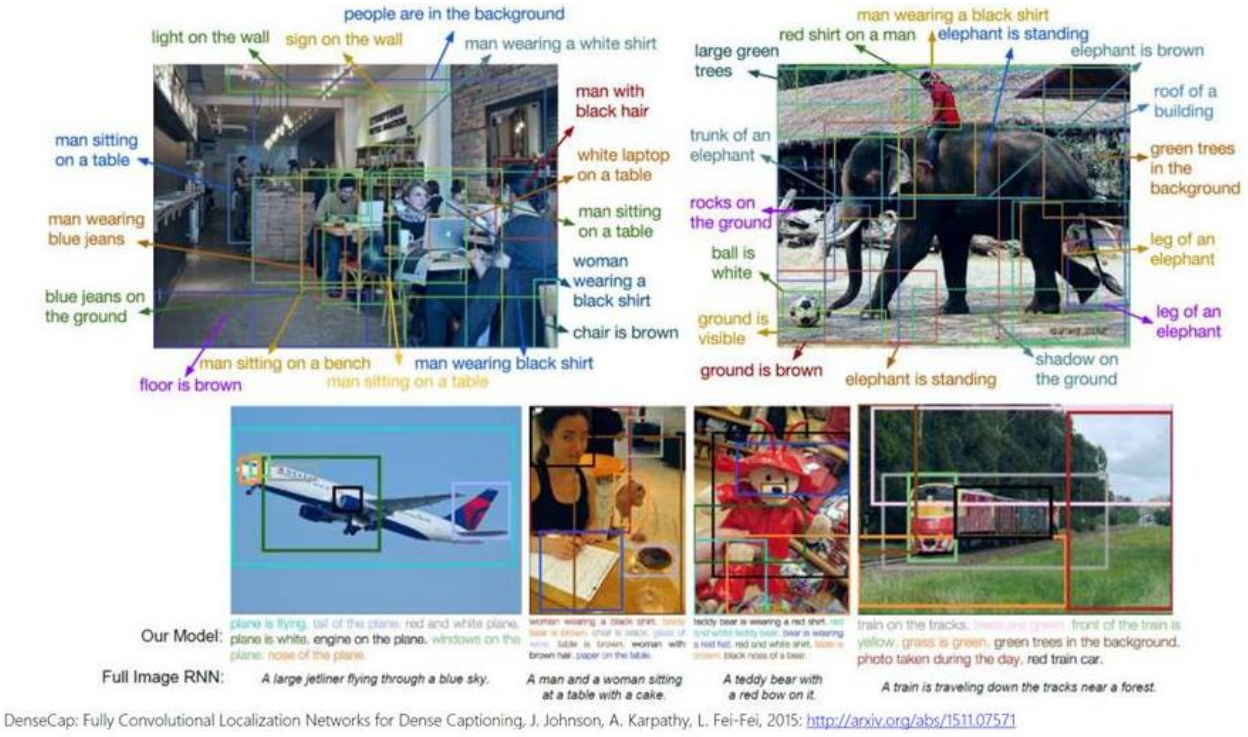
Classification d'images

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 tags qui sont ensuite exploitées dans diverses applications.

tagging automatique d'images

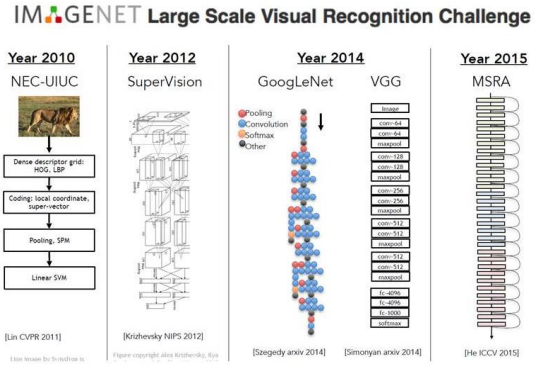
des tags aux descriptions en clair

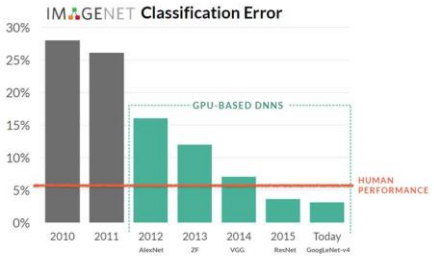
Le tagging d'images progresse chaque année. Il est maintenant possible d'identifier plusieurs objets et personnes dans une même image. Les réseaux de neurones utilisés sont de plus en plus profond (nombre de couche) et de plus en plus larges (nombre de catégories d'objets reconnus et taille des bases d'entrainement).

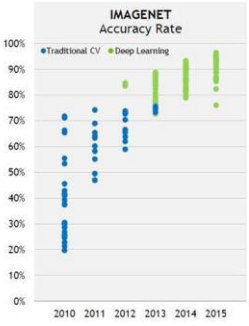


DenseCap: Fully Convolutional Localization Networks for Dense Captioning, J. Johnson, A. Karpathy, L. Fei-Fei, 2015: http://arxiv.org/abs/1511.07571

L'état de l'art progresse notamment avec le challenge **ImageNet Large Scale Visual Recognition Challenge** (ILSVRC) lancé en 2010 et renouvelé chaque année[105]. 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. Les réseaux de neurones convolutionnels actuels de reconnaissance d'images comprennent plus de 150 couches de neurones.







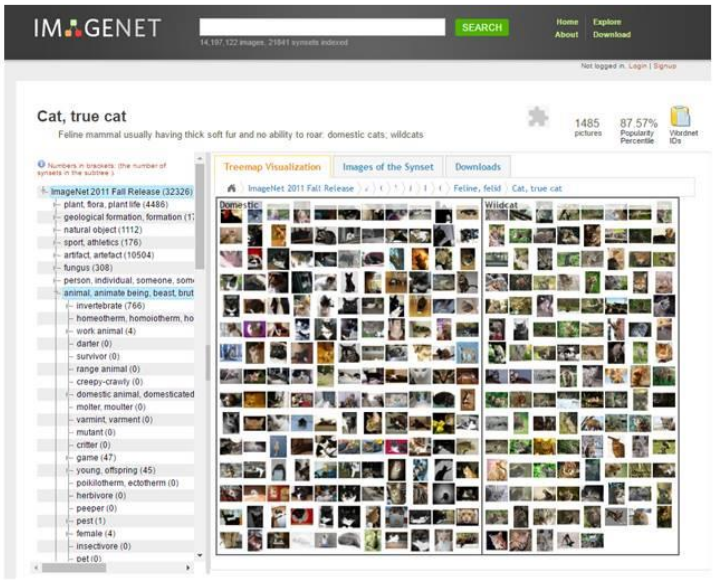
En 2016, la base de référence d'images ImageNet en comportait plus de 10 millions. Elle sert de benchmark aux solutions d'IA. Google et Facebook disposent de bases d'entrainement encore plus grandes, de plus de 100 millions d'images pour le premier et avec près de 20 000 classes d'objets différentes.

L'augmentation de la base de référence ne change rien au dimensionnement du réseau de neurones convolutionnel. Il rallonge son temps d'entrainement. L'augmentation du

[105] 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 !

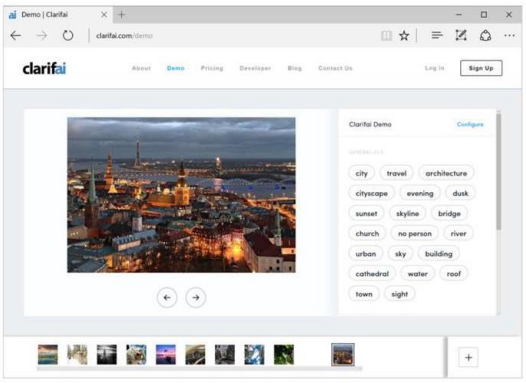
nombre de classes d'objets complexifie le réseau dans les couches finales de neurones dites « fully connected » qui font le lien entre les dernières feature maps et les classes d'objets.

base d'images de référence pour les tests de reconnaissance et classification



Voyons ce que font les startups dans le domaine :

Clarifai (2013, $72m), déjà cité, 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.



Vicarious (2010, $72m) 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%.

Cortica (2007, $38m) 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 reconnaitre une marque et modèle de voiture dans une vidéo ou un animal dans une photo (*ci-dessous*). Le tout est protégé par une centaine de brevets ! La société est originaire d'Israël.

Superfish (2006, $19,3m) développe des moteurs de recherche d'images pour les applications grand public.

Deepomatic (2014, $950K) 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. C'est une startup française !

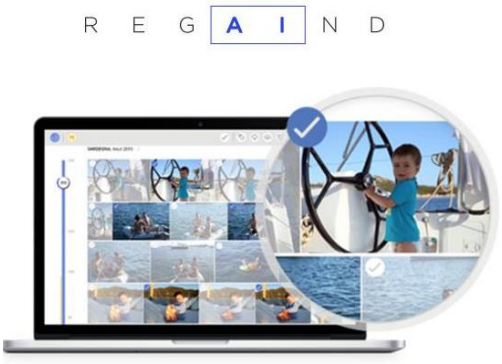
De son côté, **Facebook** sait reconnaitre un sport dans une vidéo en s'appuyant sur des réseaux neuronaux. Quant à **Google** Brain, il est capable d'identifier des chats dans des vidéos mais avec un taux d'erreurs encore très élevé, de l'ordre de 25%. La reconnaissance des visages est précise à 81,7% près (**source**). Il faut un début à tout !

Détection d'images similaires

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.

Détection de la qualité des photos

Regaind (2014, 400K€) est une startup française qui propose une solution de tri automatique de photos en cloud s'appuyant sur du machine learning et de deep learning. Elle permet de trier les photos sous un angle à la fois narratif et descriptif et de les tagger automatiquement. Elle compare 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.



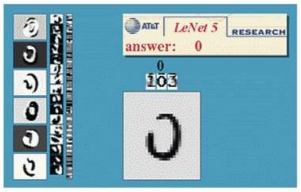
tri qualitatif automatique de photos pour créer des albums

racheté par une société US en 2017

Google Photo propose aussi une fonction équivalente.

Reconnaissance de caractères

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 !

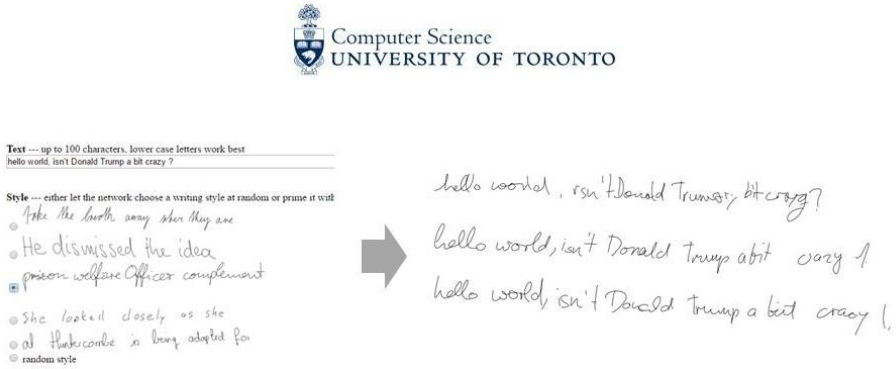
reconnaissance de caractères (OCR)

Reconnaissance d'écriture manuscrite

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**, anciennement Vision Objects, qui est basée à Nantes et qui a notamment vendu son logiciel à **Samsung**.

Nous avons aussi le californien **Captricity** (2011, $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-dessous*).



génération automatique d'écriture manuscrite

Détection d'activités

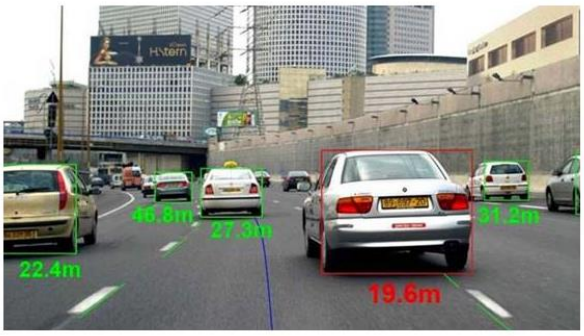
Qui est exploitée dans les systèmes de vidéo surveillance qui exploitent photos et vidéos. C'est ce que propose la startup taïwanaise **Umbo CV** (2014, $2,8M) qui gère des caméras de surveillance avec une solution logicielle fonctionnant dans le cloud.

Imagerie médicale

Pour la détection automatisée d'un grand nombre de pathologies au niveau de l'œil, de l'oreille, en dermatologie et en cancérologie. Nous verrons cela plus en détail dans la rubrique sur la santé.

Conduite assistée et autonome

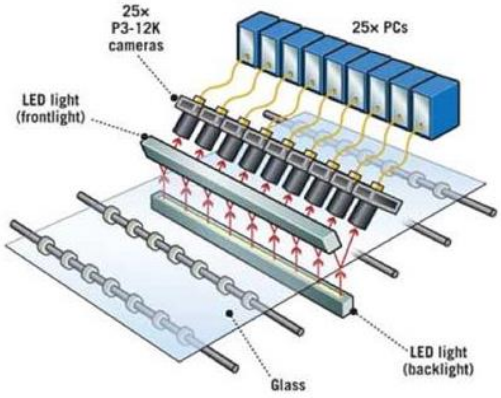
Avec par exemple les solutions de **Nvidia** et **Mobileye** qui détectent les piétons, les cyclistes, les autres véhicules, la signalisation au sol et les panneaux de signalisation. Nous le traiterons dans la rubrique dédiée aux transports.

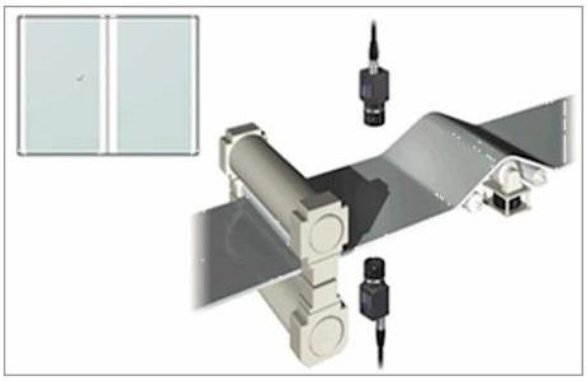


reconnaissance d'objets en temps réel pour la conduite assistée ou autonome
position, type, vitesse, risque de collision

Contrôle qualité en usine

Il est très courant et se démocratise pour vérifier la qualité des pièces et produits fabriqués en usine. 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 faible.





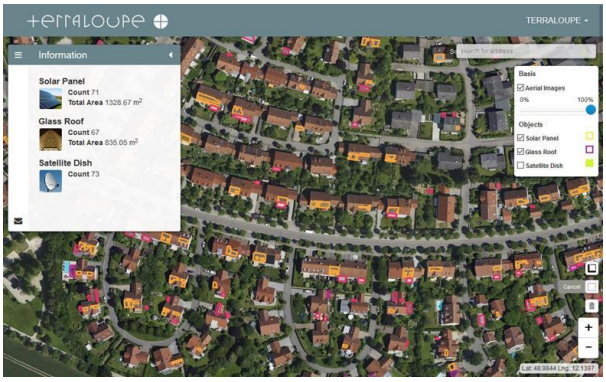
Les solutions de contrôle qualité s'appuient sur l'exploitation d'imagerie dans le visible, l'infrarouge et même les rayons X. L'imagerie peut-être complétée par d'autres types de capteurs comme ceux qui mesurent la variation de la résistance électrique de matériaux. Tous ces systèmes doivent fonctionner 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 entrainés.

Télédétection

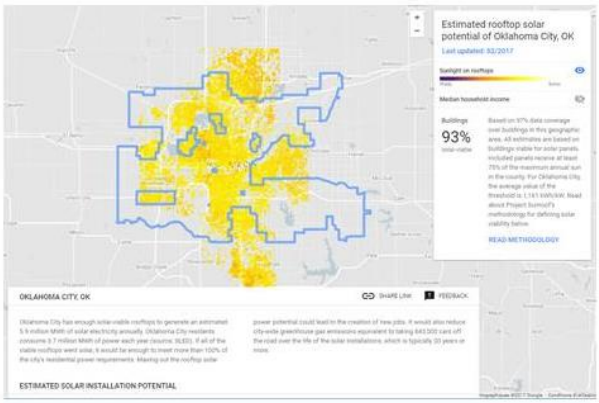
La reconnaissance d'images a aussi des applications en télédétection et imagerie satellite. Le deep learning permet de créer des solutions de recherche sémantique dans de gros volumes d'images, pour détecter des objets spécifiques comme des champs agricoles, des panneaux solaires ou des éoliennes, pour les caractériser en fonction de leur spectre lumineux, et pour analyser des variations dans le temps de ces paramètres.

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'entrainement 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, $8,28m) 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).



La startup allemande **Terraloupe** (2015, $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, $14M).



Google a même lancé son application **Sunroof** qui permet d'évaluer l'intérêt économique d'installer des panneaux solaires photovoltaïques chez soi (*ci-dessus*).

Modèles génératifs

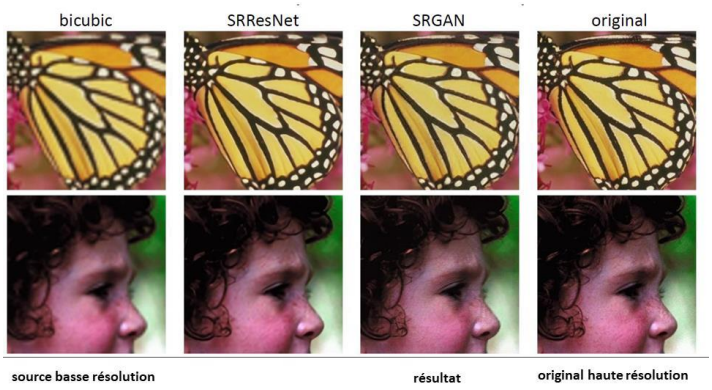
Apparus très récemment, autour de 2016, les modèles de réseaux de neurones convolutionnels génératifs, ou Generative Adversorial Networks (GAN) impressionnent par leurs capacités à prédire l'univers visuel à partir de peu d'informations. Ils complètent des images ou les transforment. Au point que l'on 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[106].

[106] Cf 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.

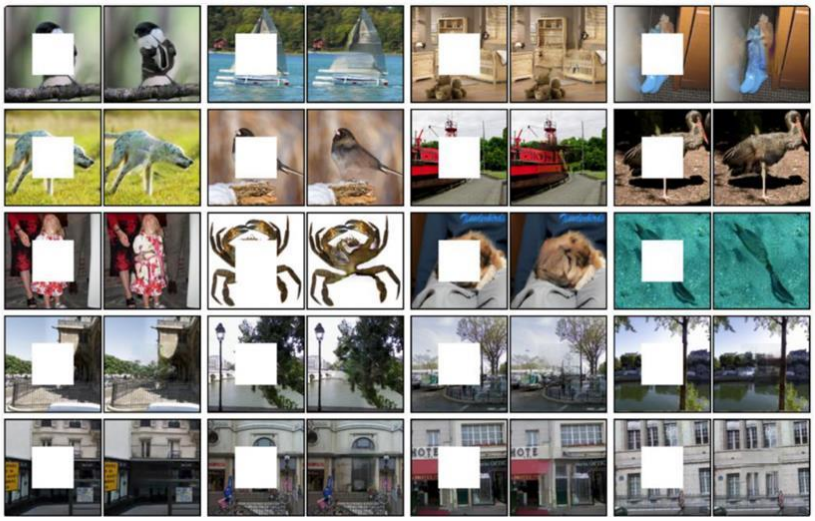
D'un point de vue pratique, ce sont des réseaux convolutionnels inversés qui génèrent un contenu à partir d'un autre contenu, éventuellement moins dense en information.

Les exemples abondent avec[107] :

- La **colorisation automatique** de photos et films en noir et blanc. Il faut évidemment de grandes bases d'entrainement.



- L'**amélioration de la résolution d'images**, *ci-dessus à droite*, qui donne des résultats étonnants[108]. Mais l'histoire ne dit pas quelle est la taille de la base d'entrainement ni si la solution fonctionne avec une grande diversité de photos. Le diable est dans les détails… que l'on ne montre pas !
- Le **complément d'images tronquées**, ci-dessous, qui donne des résultats étonnants mais imparfaits.



Pathak et al 2016

- L'**application d'un style à une photo**, souvent montée en épingle comme étant de la créativité alors qu'il s'agit d'un automatisme. Ses applications commerciales sont 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 méca-

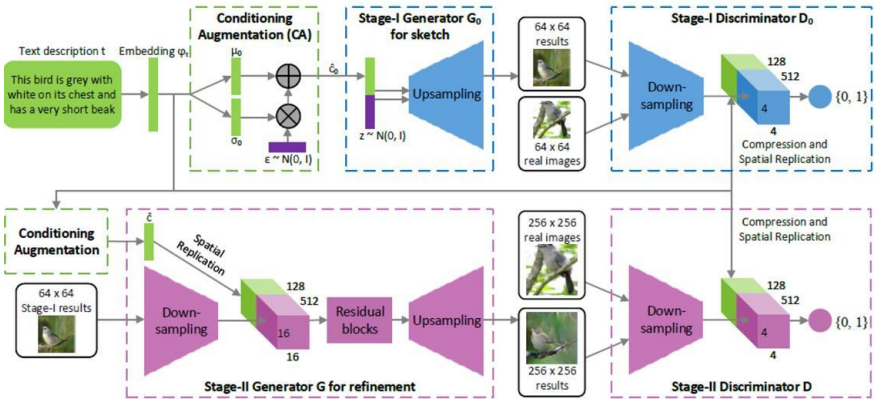
[107] Ils sont principalement issus de cette présentation : Generative Adversarial Network (GAN), de Hongsheng Li.

[108] Cf Super resolution with Generative Adversarial Networks de Boris Kovalenk.

nique qui n'est pas aussi créatif que l'on pourrait le croire[109] ! L'exemple en bas provient de Li et Wand[110].

- Encore plus fort, la **génération d'une image** à partir d'un descriptif textuel qui utilise des empilages de réseaux GAN (Generative Adversatial Networks).

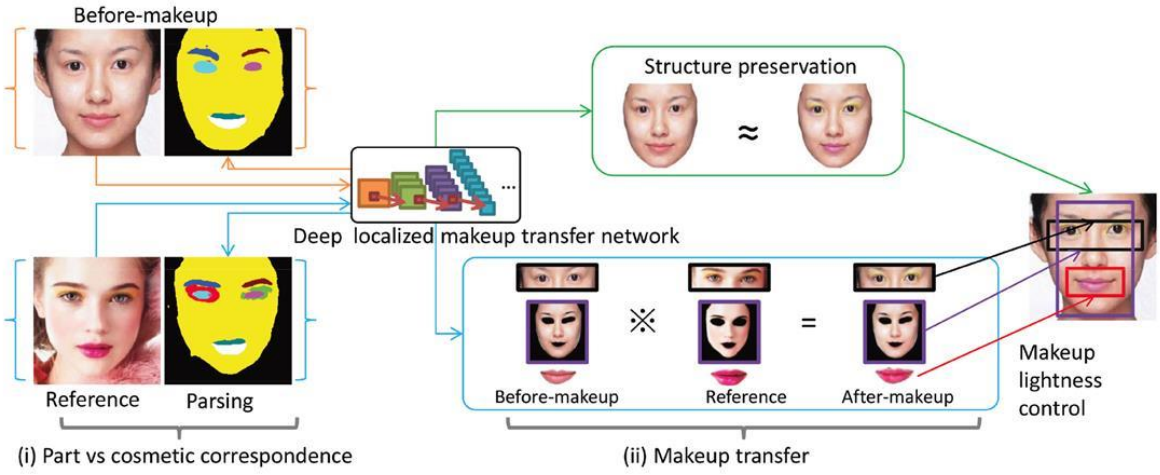
Zhang et al. 2016



[109] Cf 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.

[110] Cf StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, 2016-2017.

- L'**amélioration de selfies**, qui est proposée par **Adobe** qui utilise son IA maison Sensei, exploitant du deep learning pour améliorer de manière semi-automatique les selfies pris avec des smartphones (vidéo).
- Le **maquillage virtuel** qui passe par une analyse du visage pour le décomposer en parties auxquelles sont appliquées ensuite divers maquillages[111].



- 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.

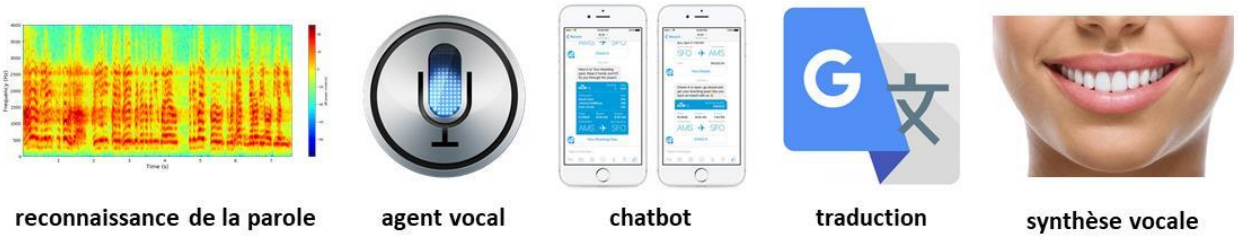
Nombre des exemples ci-dessus sont issus de laboratoires de recherche et sont par nature expérimentaux et surtout imparfaits. Certains des procédés se retrouvent néanmoins intégrés dans des startups comme pour Loom.ai.

Langage

Le traitement du langage est le second plus grand domaine d'applications de l'IA avec la vision artificielle. Il comprend de nombreuses composantes 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 synthèse de textes ou de paroles.

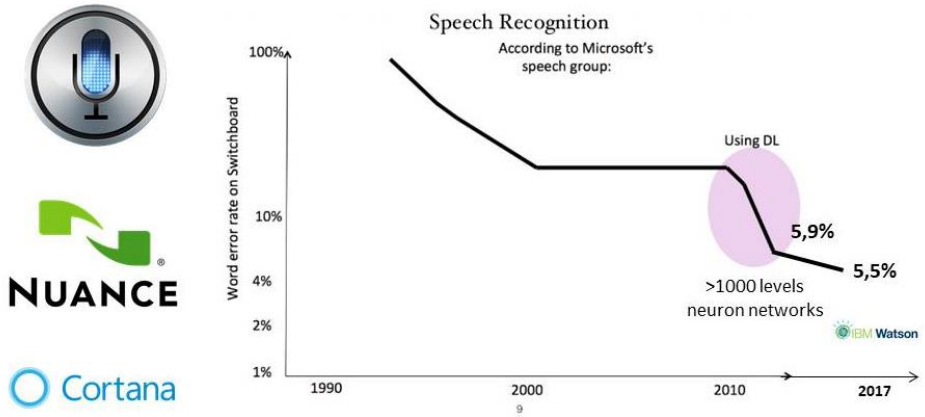
Le domaine exploite surtout le deep learning, surtout pour la reconnaissance de la parole. Le deep learning est aussi de plus en plus utilisé pour extraire les connaissances des textes et pour la traduction automatique.

[111] Cf Makeup Like a Superstar: Deep Localized Makeup Transfer Network de Si Liu, Xinyu Ou, Ruihe Qian, WeiWang et Xiaochun Cao, 2016.



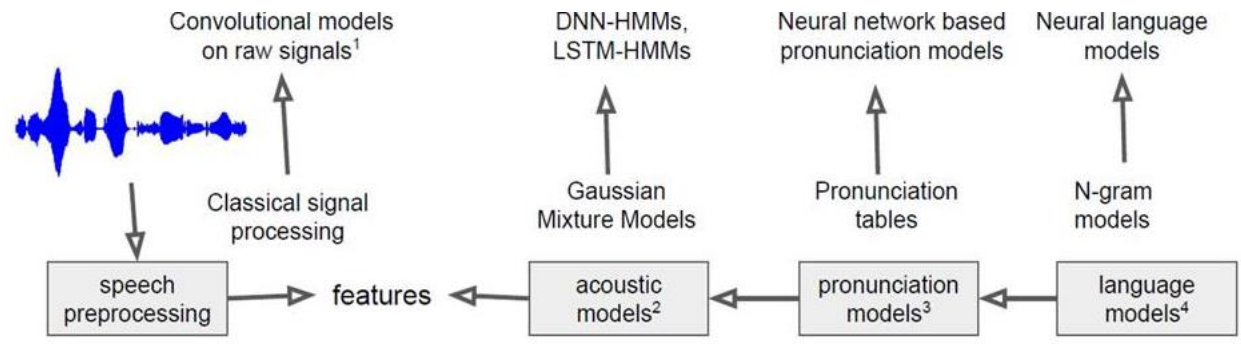
Reconnaissance de la parole

La reconnaissance de la parole s'appuyait au départ sur des techniques statistiques et notamment bayésiennes. Elle a fait des progrès continus grâce à l'intégration de techniques différentes telles que le deep learning, le big data, les réseaux neuronaux et des modèles de Markov à base de statistiques.



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'à1000 ! Le taux d'erreurs de la reconnaissance est maintenant inférieur à celui de la compréhension humaine, quc ce soit chez **Microsoft** et **IBM**.

Les solutions de reconnaissance vocale ont encore souvent besoin d'accéder à des bases de données de référence, surtout s'il fonctionne sans apprentissage de la voix de l'utilisateur. Cela nécessite un aller et retour avec les serveurs du service, ce qui est fréquent avec les smartphones et avec des services comme SIRI d'Apple. D'où l'intérêt de la 4G et de son débit comme de son faible temps de latence pour les allers et retours avec les serveurs.



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.

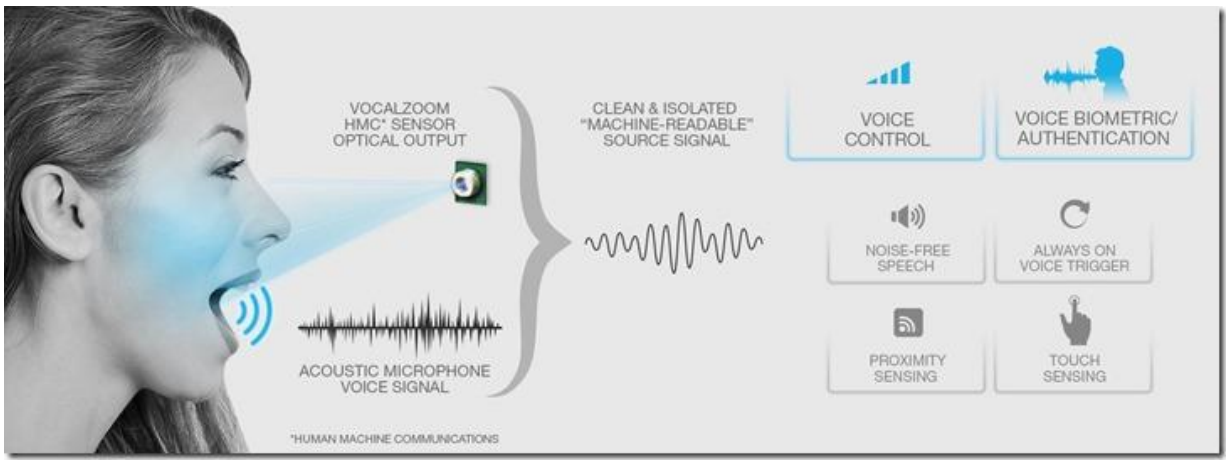
Pour en savoir plus, voir cet historique de la recherche en reconnaissance de la parole : **Survey of Technical Progress in Speech Recognition by Machine over Few Years of Research** parue en 2015. Ce sujet intègre de nombreuses branches du savoir issu de plusieurs décennies de recherches dans l'IA.

On est encore **loin de la solution parfaite**, notamment parce que les logiciels manquent d'informations sur le contexte des conversations[112]. 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. D'où l'intérêt de la récente publication en open source de la solution **Deep Speech 2** de **Baidu** qui fonctionne en anglais et en chinois[113].

Le taux d'erreur est particulièrement élevé s'il y a du bruit ambiant, 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** 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.

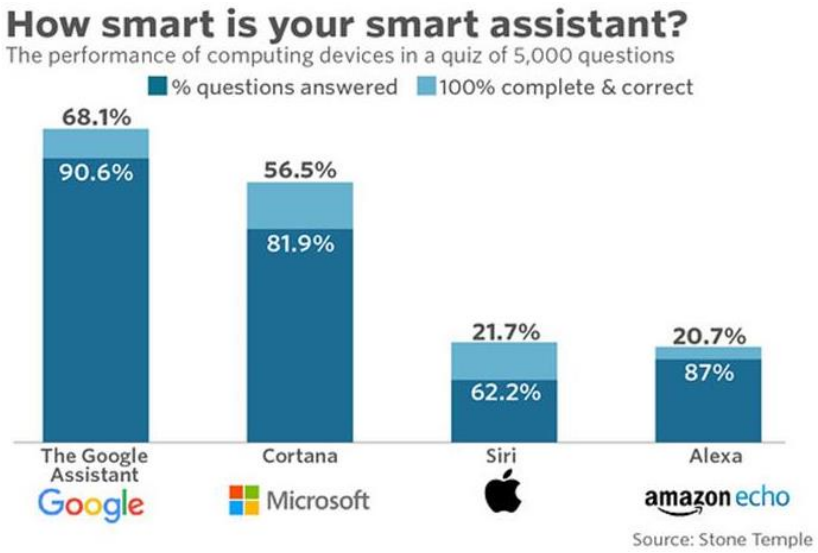
[112] Voir aussi **Why our crazy smart AI still sucks in transcribing speech** paru dans Wired en avril 2016.

[113] Cf 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 entrainé avec 12 000 heures de conversations. Les versions les plus récentes ont été entrainées avec plus de 100 000 heures de conversations en environnement bruyant.



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 (OK **Google**, **Microsoft** Cortana, **Apple** Siri, **Amazon** Alexa, **Samsung** Bixby qui est probablement originaire de **Viv Labs**[114]).

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. 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.



Ces différents assistants se distinguent dans leur capacité à répondre à des questions diverses. Un récent benchmark met les assistants de Google et Microsoft devant ceux d'Apple et Amazon (*ci-dessus*[115]).

---

[114] 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.

[115] Cf « Alexa and Cortana Will Talk to Each Other Say Amazon and Microsoft », dans Voicebot.ai, août 2017.

Ce qui explique peut-être l'annonce fin août 2016 d'un partenariat entre Microsoft et Amazon qui vont faire en sorte que Cortana puisse dialoguer avec Alexa et réciproquement, rendant ainsi, via la voix, leurs bibliothèques applicatives de services compatibles.

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 reconnaitre ses erreurs, etc.

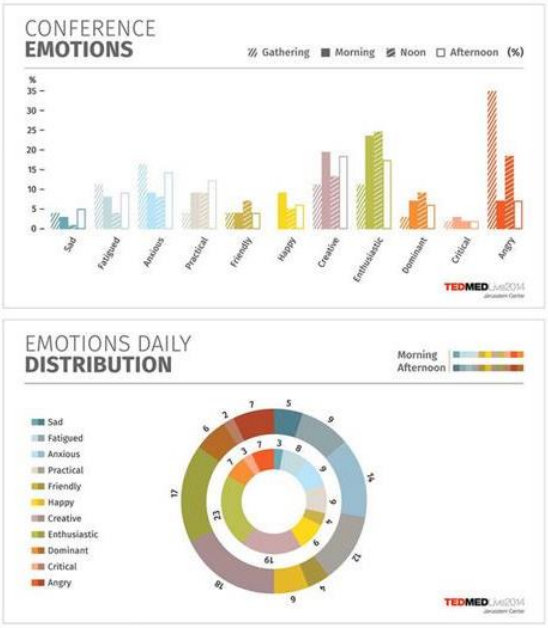
L'américain **Nuance,** qui dépasse $2B de chiffre d'affaire, vend sa solution un peu partout en OEM. Apple a fait l'acquisition de la start-up **VocaliQ** en 2015 et **Sensory** fait avancer l'état de l'art de manière indépendante depuis plus de 20 ans. Le couteau Suisse IBM Watson peut aussi servir 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, $251m) est une startup basée à 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 !

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 le français **BatVoice** 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, l'israélien **BeyondVerbal** (2012, $11m) qui commercialise de la propriété intellectuelle issue de longues années de recherche dans le domaine.

Dans le même ordre d'idée, **Cogito** (2007, $22,5m) analyse les appels dans les centres d'appels pour donner 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) et **Cognitive Matchbox** (2016) proposent chacune une solution de routage d'appels optimisée aux centres d'appels qui analyse la personnalité et les émotions des clients pour les orienter vers le meilleur agent. 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.

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).

Génération de parole

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.

Des startups abordent aussi sur ce marché comme la canadienne **Lyrebird** (2017) qui proposera sa solution aux développeurs sous forme d'API en cloud, exploitant des serveurs à base de GPU Nvidia et qui était encore en bêta en septembre 2017. 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.[116]

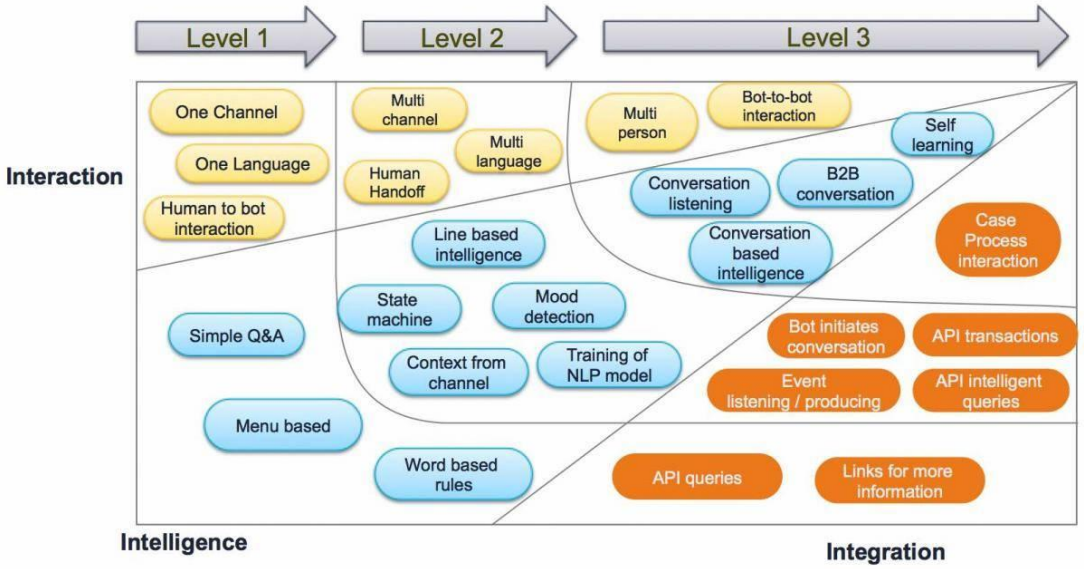
[116] Cf leurs démonstrations avec les voix de Donald Trump et Barack Obama : https://lyrebird.ai/demo.

Chatbots

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, …).

Les chatbots visent à automatiser le service client en ligne dans les sites de e-commerce, services financiers ou autres. 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 dans la pratique et pas forcément appréciés des utilisateurs. Ces chatbots peuvent avoir une interface vocale comme avec SIRI, Cortana et Amazon Alexa.

Il est assez difficile d’évaluer la maitrise technologique des différentes sociétés de ce secteur. Elles utilisent un patchwork de différentes APIs et outils de deep learning plus ou moins packagés[117]. 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[118].



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 trai-

[117] 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.

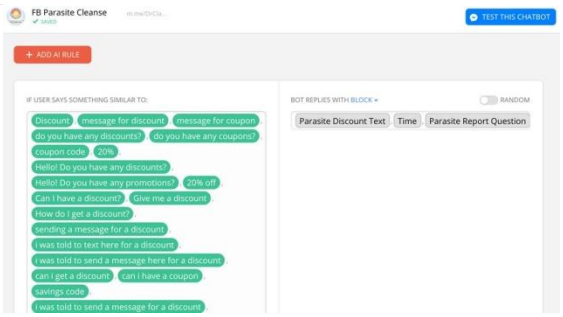
[118] Cf L’arnaque chatbots durera-t-elle encore longtemps ? de Par Thomas Gouritin, octbre 2017.

tement du langage, à des modèles prédictifs et en tenant aussi compte de l'humeur du client. On appelle cela un chatbot de niveau 3[119].

Dans tous les cas de figure, un bon chatbot doit être alimenté par des sources d'information diverses :

- 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"]
        },
```



- 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.

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[120]. 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.

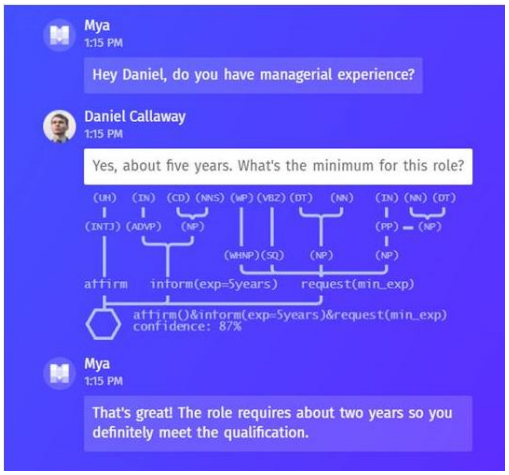
---

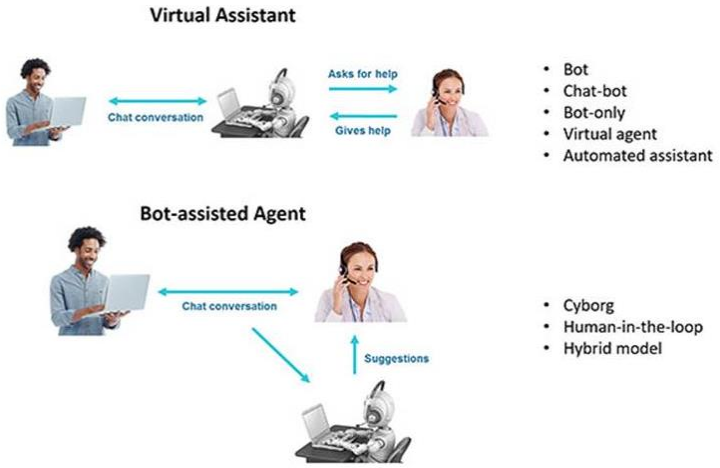
[119] Le schéma au-dessus qui décrit les caractéristiques de trois niveaux de chatbots provient de How can Chatbots meet expectations? Introducing the Bot Maturity Model, Léon Smiers, Oracle, avril 2017.

[120] Cf http://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3.

Les chatbots sont de trois types différents du côté des interactions :

- 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.



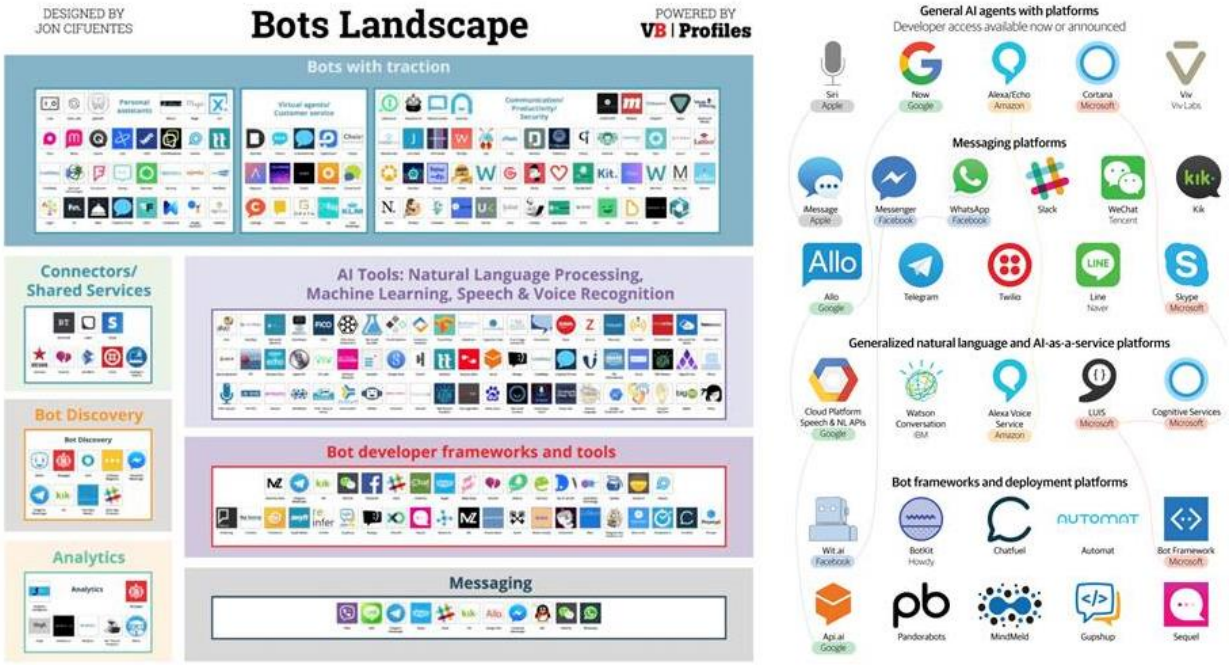


L'offre peut être segmentée avec des chatbots généralistes, des chatbots spécialisés dans des domaines précis (ecommerce, recrutement, …) et des outils de création de chatbots[121] et des plateformes d'accueil de chatbots comme Facebook Messenger ou Slack.

Dans la pratique, 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 !

[121] Cf 25 Chatbot Platforms: A Comparative Table par Olga Davydova, mai 2017, qui recense et compare 25 outils de création chatbots.



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 :

- Du côté des **chatbots vocaux**, Amazon Alexa semble nettement en avance. Cet outil est surtout utilisé dans la maison connectée. Il est suivi de SIRI et de Google Assistan ;
- Du côté des **chatbots textuels**, 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 :

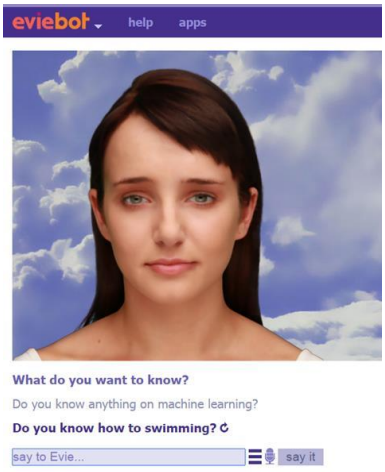
- **Semantic Machines** (2015, $12,38M) est une startup de Boston et Berkeley qui propose des chatbots pouvant être intégrés dans toutes sortes d'usages, b2b et b2c. L'équipe fondatrice comprend des anciens de Siri et Google Now. La solution intègre la reconnaissance et la synthèse de la parole.
- **Talla** ($4M) 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.
- Puis les 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, $2,7m) qui est notamment déployée chez Sony, **Niki.ai** (NC) est une startup indienne qui se focalise dans les services (transports, voyage aérien, santé) et **ReplyYes** (2015, $3,5m), une startup de Seattle, qui associe machine learning et opé-

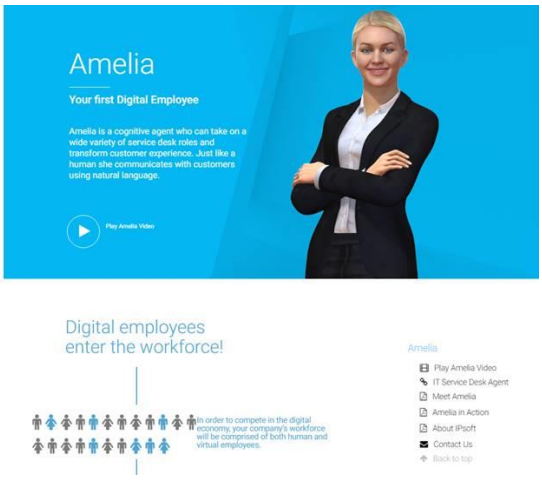
rateurs humains avec deux spinoffs, l'une qui commercialise des disques vinyles (The Edit) et l'autres, des BD (Origin Bound). The Edit aurait vendu $1m de vinyles en huit mois.

- **TARA** (NC) est une startup de San Francisco qui propose un robot conversationnel de gestion du recrutement de freelances.
- **Do You Dream Up** (2009) est une startup française qui 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.

Passons à quelques solutions de chatbots plus originales :

- **IPSoft** (1998) est une startup assez ancienne de New York qui propose son chatbot Amelia qui est positionné sur le helpdesk IT. Il exploite un agent de détection d'émotion des utilisateurs. La startup française **Gorgias** (2015, $1,6M) est aussi positionnée sur l'automatisation du helpdesk IT. Mais son outil aide les conseillers de support à être plus efficaces, sans les remplacer.
- **Existor** (1988) est une startup anglaise créatrice d'agents conversationnels comme 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-dessous*). 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 !





- **Snips** (2013, $19,3m) est une startup française créée par Rand Hindi, avec un dernier tour de financement de $13M en juin 2017 par Korelya, le fonds d'investissement de Fleur Pellerin, accompagné par MAI Avenir. La société propose un assistant vocal pour applications mobiles. Sa particularité est de fonctionner en mode autonome sans nécessiter un aller et retour avec un serveur. Au passage, 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 inté-

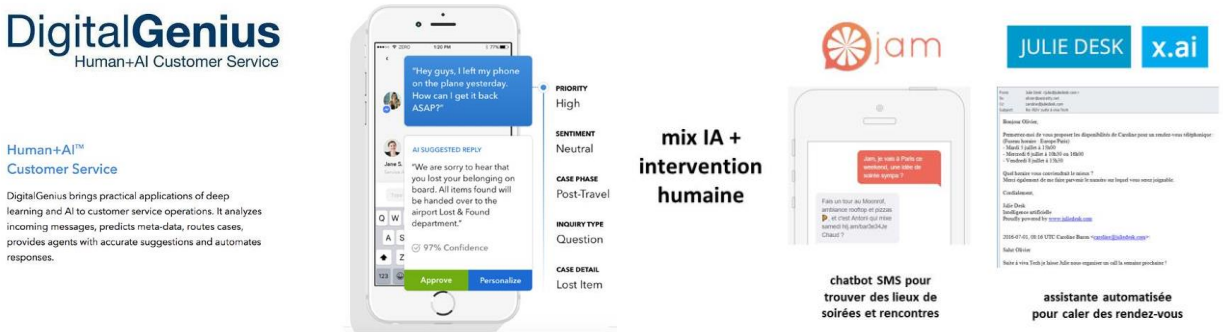
grable dans des applications existantes. La solution est commercialisée à un prix fixe qui ne dépend pas du nombre d'utilisateurs.

- La startup américano-russe **Luka** (2014, $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. Dans la pratique, il sert surtout à choisir des restaurants (*ci-dessus à droite*). 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.
- **Publicis** a développé une application de recommandation de maquillage à base de chatbot pour Sephora[122]. Cela relève encore d'une approche de service sur mesure, pas de la création d'un produit.
- Le français **Hubware** (2016) 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.

Passons aux chatbots associant automatisation et intervention humaine :

- **Curious.ai** (2013, $7,35m) est une startup américaine qui commercialise Digital-Genius 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 entrainé avec des transcriptions d'appels réels au service client.

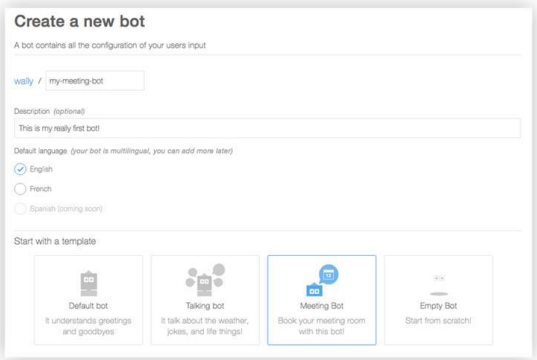
[122] Source : keynote de Microsoft AI en septembre 2017, https://myignite.microsoft.com/sessions/56555.

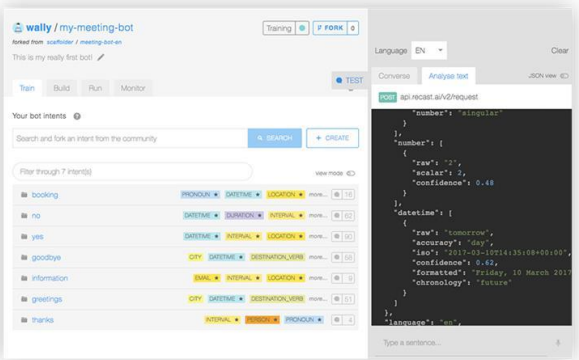


- **Jam** (2012, $1,46m) et **Julie Desk** (2014, $3,7m) sont deux startups françaises qui proposent des solutions d'agents conversationnels intégrant également une assistance humaine. Jam est un agent SMS qui permet à des étudiants d'organiser leurs soirées. De vrais gens répondent à la main lorsque le système ne peut pas le faire. Les techniques utilisées par Hello Jam ne sont pas bien documentées. Elles ne semblent pas faire appel à du deep learning sophistiqué. Julie Desk 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, $44,3M).

Et enfin, voici quelques outils de création de chatbots exploitables par les entreprises et agences spécialisées de création de chatbots :

- **recast.ai** (2015, $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[123] comme un acteur clé de la création d'agents conversationnels.





- Le BotMaker de **Viseo** (1999, $2M) est un outil de création de chatbots textes et vocaux en open source doté d'une interface graphique interactive capable notam-

[123] Ici : http://aiplaybook.a16z.com/docs/guides/nlp.

ment 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.

- La startup française **Opla.ai** (2015), 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, $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), une startup d'Oakland (Californie) qui propose une plateforme de chatbot en ligne, open source et multi-lingue. 285 000 chatbots avaient été générés avec en date d'août 2017. Ils sont intégrables dans divers environnements de messagerie instantanés tels que Slack et Whatsapp.
- **Viv Labs** (2012, $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 GAFAM, notamment Messenger 2.1 chez **Facebook** qui permet de développer son propre chatbot, les outils développeurs de Cortana chez **Microsoft**, ceux de **Google** et enfin, **Amazon** Alexa, qui est très utilisé dans le domaine des objets connectés pour les rendre pilotables par la voix.

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.

Traduction automatique

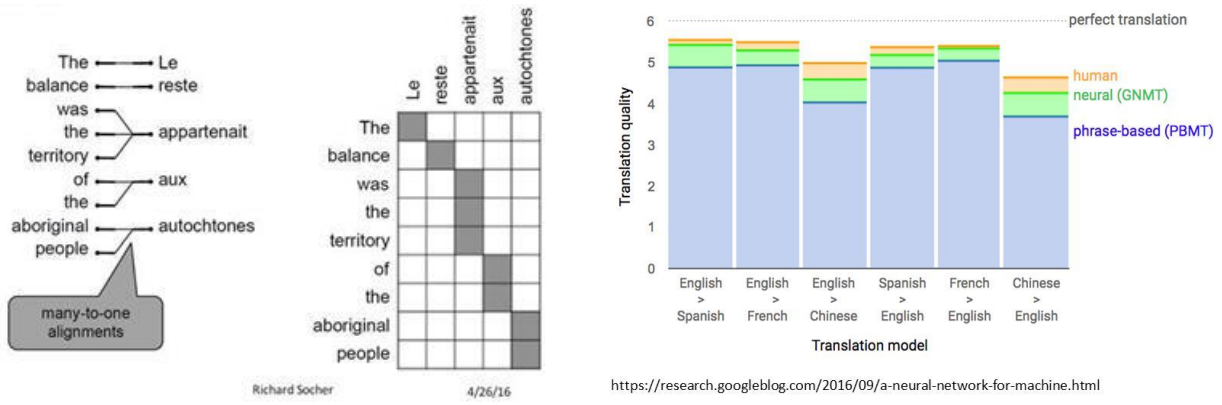
La traduction automatique s'est longtemps appuyée sur des méthodes statistiques avec énormément de bidouillage manuel.

Le deep learning a fait son apparition dans le domaine relativement récemment. Il exploite des réseaux de neurones récurrents (RNN), leur variante à mémoire (LSTM : Long Short Term Memory) et de nombreuses autres déclinaisons[124]. Ce champ d'application s'appelle en américain le Neural MT pour Neural Machine Translation. Ce courant est devenu dominant très récemment, 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. La traduction à base d'IA est en-

[124] Cf 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.

core 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[125].



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[126].

L'un des leaders mondiaux 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.

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. Mais elles n'ont pas la même fonctionnalité : Google Translate traduit le texte photographié dans des images tandis que Cortana fait du speech-to-speech. Google Translate 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.

Extraction de données

Le traitement du langage a d'autres applications diverses consistant à exploiter les données qualitatives 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

[125] Visualisable ici : https://www.youtube.com/watch?v=Nu-nlQqFCKg.

[126] La méthode est documentée dans cette présentation « Deep Learning in Speech Synthesis » de Heiga Zen de Google qui date de 2013 : https://static.googleusercontent.com/media/research.google.com/fr//pubs/archive/41539.pdf.

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.

L'exploitation de réseaux de neurones RNN et de LSTM permet aussi d'améliorer la capacité à détecter les sentiments dans des textes, de manière plus fine qu'avec une simple analyse syntaxique classique[127]. Ces techniques servent même à détecter les discours de haine sur Internet[128].

De nombreuses startups opèrent dans ce secteur. L'anglais **Wrapidity** (2015) 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) début 2017. **DefinedCrowd** (**vidéo** avec son ukulélé de circonstance) capte les données vocales ou textuelles et les exploite, notamment pour de l'analyse de sentiments. **Weotta** (2011) met en œuvre ce genre de technique dans son application WeottaGo, une application de recommandation mobile.

Le français **Heuritech** (2013) 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. 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).

Proxem (2007, 1m€) est une autre startup française qui propose 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 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.

La startup française **Keluro** (2014) 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 en bêta depuis septembre 2016.

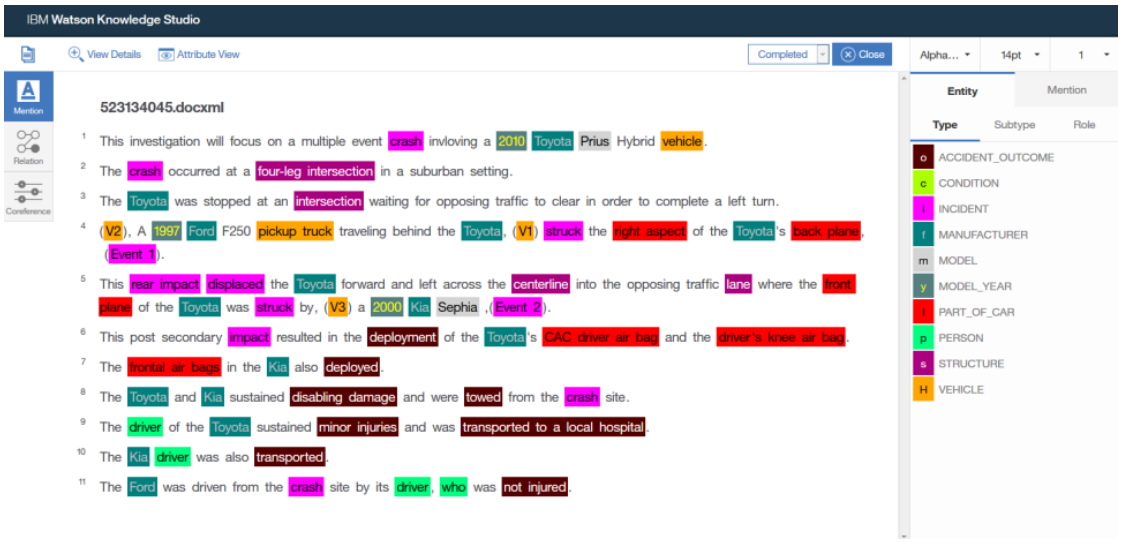
SkipFlag Rover exploite les données circulant dans les entreprises notamment dans les outils collaboratifs comme Slack pour créer des bases de connaisances capables de répondre aux questions clés. Encore une application exploitant l'analyse du langage.

---

[127] Cf OpenAI sets benchmark for sentiment analysis using an efficient mLSTM, avril 2017.

[128] Cf Internet researchers harness the power of algorithms to find hate speech, octobre 2017.

DeepQA, l'une des briques d'**IBM Watson**, permet de créer des agents conversationnels. Elle permet d'extraire des règles de textes, comme des documents scientifiques. Cela permet d'alimenter en retour des moteurs de règles de systèmes experts.



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.

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.

En France, **Antidot** (1999, $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, $5,33m) est un fournisseur français 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, $422K), une startup d'Oslo 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. 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).

Narrative Science (2010, $40,4m) 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. 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.

De nombreuses startups sont positionnées sur ce secteur, comme **Arria** qui vise les marchés financiers, des utilities, de la santé et du marketing, les français **LabSense** et **Syllabs**, les Allemands **Retresco** (2008) qui produit automatiquement des comptes-rendus de compétitions sportives et **Textomatic** (2010, $40M) ou **Automated Insights** (2007, $10,8M).

Robotique

La robotique est un domaine à part entière qui tire de plus en plus partie des briques techniques de l'IA.

La notion de robot est très ancienne et 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.

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 humanoides. Cette catégorie de robots évo-

lue 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), 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.

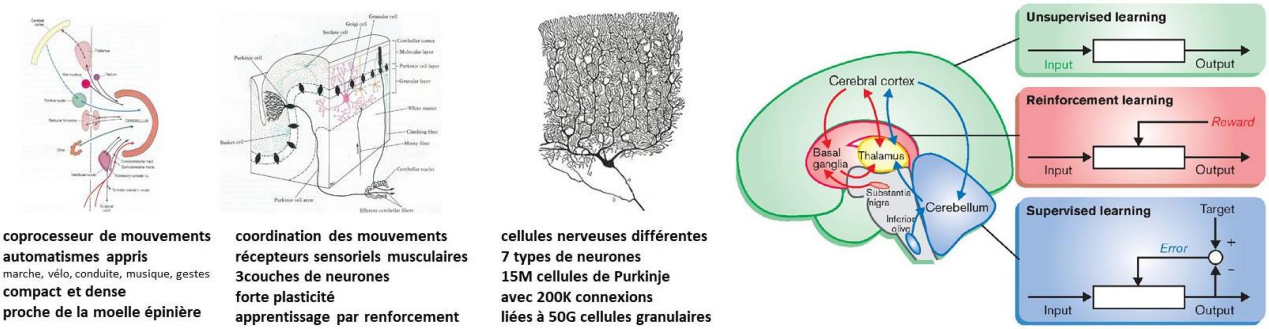
Un robot est dans la pratique un agent composé de très nombreux agents qui doivent être bien coordonnés. Il doit accomplir des tâches avec plus ou moins de degrés de liberté et d'initiative. Il doit pouvoir s'adapter à son environnement et gérer les imprévus. Et enfin, il doit respecter les fameuses lois de la robotique de l'écrivain Isaac Asimov issues de "I, Robot" (1950)[129].

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 ?

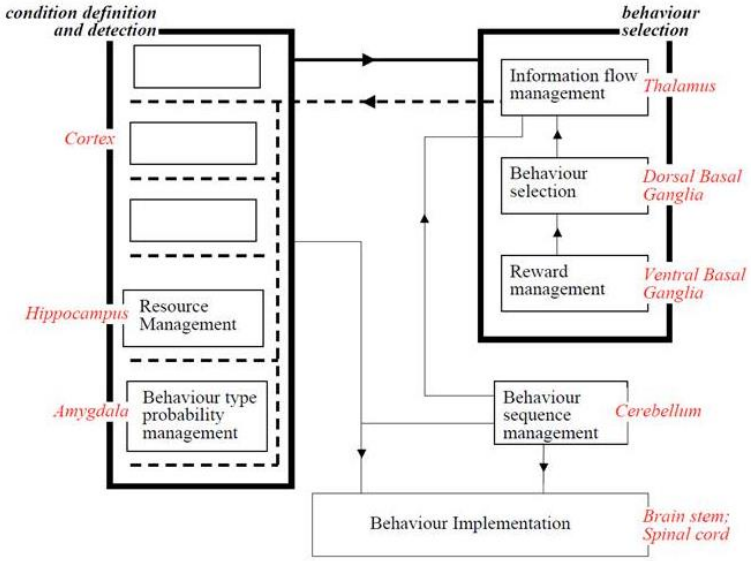
En robotique comme dans le reste de l'IA, le biomimétisme est aussi une source d'inspiration. L'articulation du fonctionnement entre le cortex cérébral et le cervelet fait l'objet de nombreuses études sur la coordination des mouvements. Le cervelet joue le rôle de coprocesseur de mouvements appris pour le cortex. Il fonctionne sur un mode supervisé avec un apprentissage progressif, comme pour celui de la marche pour les enfants en bas âge.

[129] 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.



Les neurones du cervelet sont organisés différemment de ceux du cortex. Le cortex comprend 16 milliards de neurones alors que le cervelet en contient au moins 56 milliards. Ils qui sont plus denses et plus connectés entre eux que dans le cortex. Par contre, alors que le cortex comprend de très nombreuses cellules gliales qui régulent leur activité et la transmission des influx nerveux des axones reliant les neurones entre eux, le cervelet en comprend beaucoup moins. Qui plus est, le cervelet est proche de la moëlle épinière. Tout est fait pour lui permettre de fonctionner en temps réel !

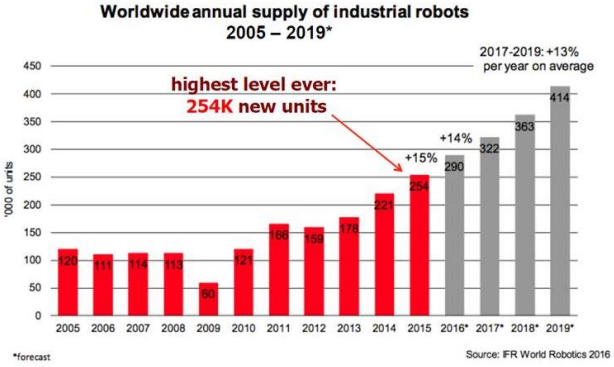
La position du cervelet dans la boite crânienne empêche d'en mesurer l'activité comme on le fait avec des capteurs d'électroencéphalogrammes autour du cortex. Sa densité est telle que même Neuralink n'envisage pas d'y placer des électrodes pour nous apprendre instantanément à bien danser ou pratiquer tel ou tel sport. Le cervelet conserve donc un caractère plus mystérieux que le cortex. Et il est encore plus difficile à imiter que le cortex. Avec la qualité de nos sens, cela explique indirectement pourquoi la robotique, surtout humanoïde, progresse plutôt lentement alors que la robotique industrielle et non humanoïde constitue l'essentiel de ce marché.

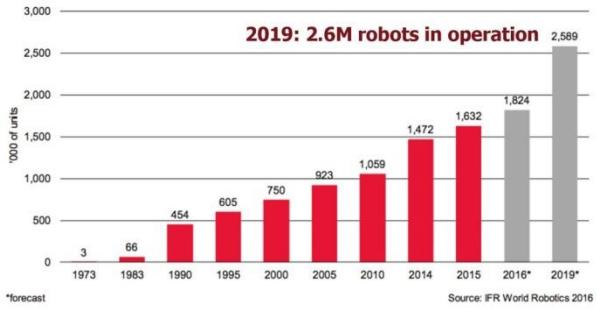


Ainsi, selon **IDC**[130], le marché de la robotique mondial était de $71B en 2015 et devrait atteindre $135B en 2019, générant une croissance annuelle de 17%. Le marché des services en robotique était estimé de son côté aux alentours de $9B à $11B en

[130] Source : http://fortune.com/2016/02/24/robotics-market-multi-billion-boom/.

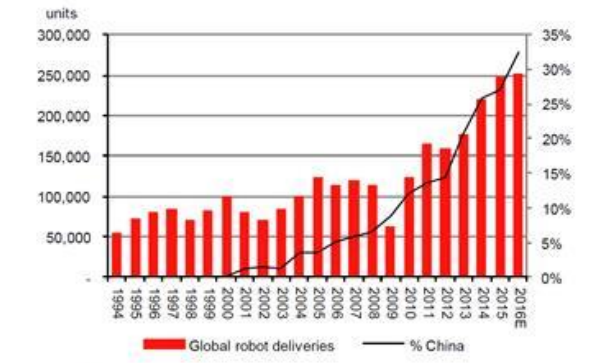
2015 selon les sources. Il s'est vendu presque 300 000 robots industriels en 2016 et c'est la Chine qui capte la plus grande partie de ce marché, tant en production qu'en installations.





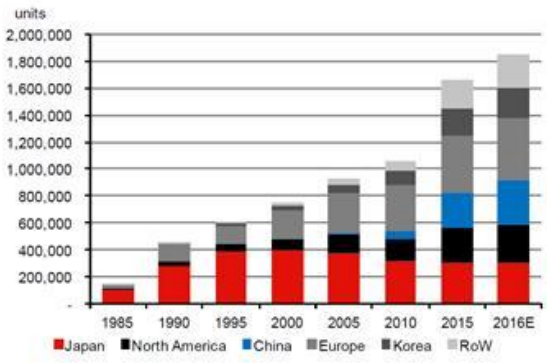
Les gros consommateurs de robots sont sans surprise les grands pays industriels : la Chine, la Corée du Sud, le Japon, les USA, l'Allemagne et l'Italie, qui est devant la France. La base installée des robots industriels serait d'environ 2 millions d'unité à la mi 2017.

Fig 16 China robot demand as a % of total



Source: IFR, Macquarie Research, September 2016

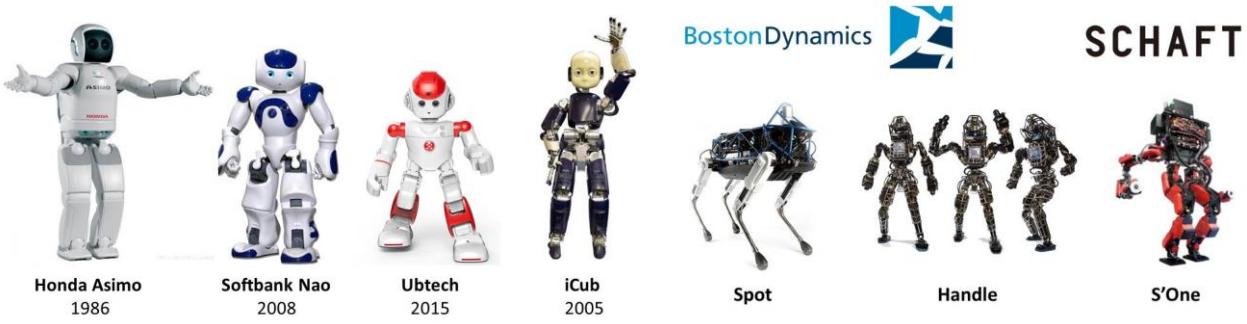
Fig 17 China has driven 36% of the installed base growth between 2010 and 2016E amongst key



Source: IFR, Macquarie Research, September 2016

Le marché est surtout concentré sur les robots industriels mais ce sont les robots humanoides 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.

Le robot humanoide le plus avancé du côté de sa capacité à se mouvoir est probablemen 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.



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 humanoides 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. Dans la pratique, 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 et Handle (*ci-dessus, à droite*), capables de rouler avec habileté et de déplacer des paquets dans des étagères. 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.

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**, un autre robot français, créé à Lyon et animé par les logiciels d'un autre Lyonnais, la startup **Hoomano**. Il y a aussi ceux d'autres startups françaises, **Partnering Robotics** (2007) et

Wyca Robotics ou le Japonais **Robosoft Kompaï** (1996, $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.

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 le français **Wandercraft** (2013, $19,4M), qui vient de démontrer en septembre 2017 son premier prototype opérationnel, ou le japonais **Cyberdyne**, et qui remonte parfois au cortex moteur dans les lobes bipariétaux comme pour le projet d'exosquelette à quatre membres du laboratoire grenoblois **Clinatec** qui est destiné aux tétraplégiques.

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 l'Américain **BrainRobotics** et l'exosquelette lombaire Atlas du français **Japet Medical Devices**.

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**[131], **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.

[131] 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. Cf https://vimeo.com/170005575.

Roomba (2002) iRobot

Friendly Robotics (2000)

AI.mergence (2017)

Intellibot (2015) Taski

Husqvarna (2015)

EOS Innovation (2010)

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) et que j'avais découverte à l'occasion du CES 2017.

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**. 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.

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** vu 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 qui** ont été acquis par Amazon pour $775m en 2012 (*ci-dessous à gauche*).

FuelMatics (2008)

Dans les transports, les robots de **FuelMatics** 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 octobre 2018 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

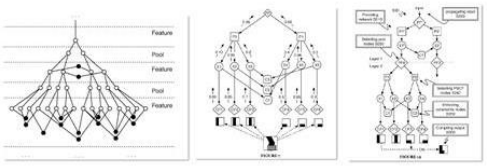
Les drones militaires savent gérer leur vol de manière autonome mais sont pilotés à distance, notamment pour les ordres d'observations ou de lancements de missiles. Il en va de même pour les drones sous-marins qui sont téléguidés.

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, $122M) et **Kindred** (2014). Des startups comme l'Américaine **Neurala** (2006, $14M) 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.

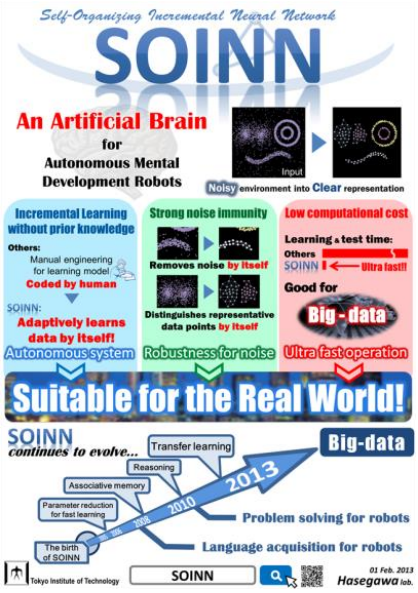
vicarious

réseaux corticaux



KINDRED

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 légo (source) ou de résolution du Rubik's Cube en une demi seconde (vidéo).

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[132].

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.

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.

Heureusement, toutes ces startups ne font pas appel à de l'IA même si un trop grand nombre s'en vante ! Et 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).

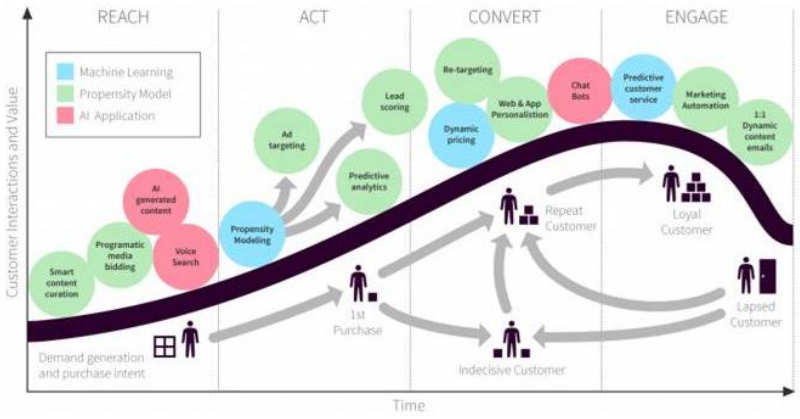
[132] Comme avec **iSee Automation**, une startup issue du MIT qui est financée dans le cadre du fonds de deep techs **The Engine** lancé par ce dernier et doté de $200M (source).

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.

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 !



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. Bref, la panoplie des outils d'IA en marketing et vente est vaste, surtout pour les sites de vente en ligne.

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

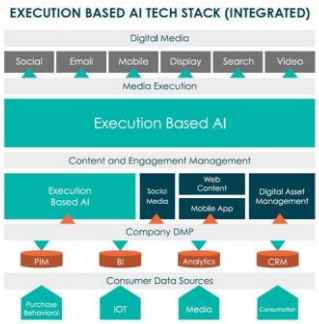
recommandation sociale : sur la confiance

Une bonne part des startups de ce secteur proposent des solutions généralistes intégrant plusieurs outils et accédant à des sources de données externes. Lucy d'**Equals3** (2015) s'appuie sur Watson pour segmenter les clients, définir ses messages et optimiser son média planning.

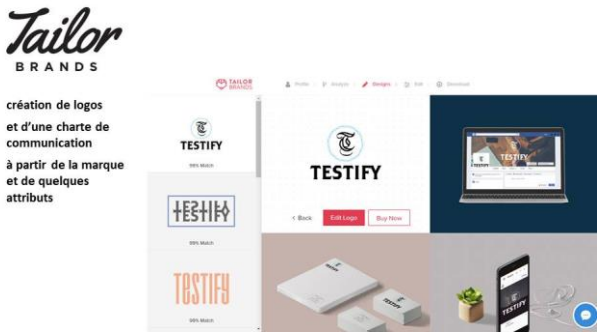
Albert de la startup **AdGorithms** (2010, IPO en 2015) intègre des outils de segmentation d'audience, d'achats médias, d'optimisation plan média cross-channel, de tests et optimisation et d'analytics. Dans la pratique, c'est une grosse boite noire exploitant de nombreuses sources de données. Albert et Lucy s'appuient tous les deux sur des briques d'IBM Watson.

L'optimisation des messages et contenus est aussi le domaine des startups américaines assez bien financées que sont **Captora** (2012, $27m) et **Persado** (2012, $66m).

La planification des messages et des médias s'appuie sur la gestion et l'analyse des données issues des médias sociaux comme avec **Meshfire** (2012, $350K), **Cortex** qui prédit la réaction des Internautes aux contenus (2014, $500K) et **SimpleReach** (2010, $24,2m).



Il existe des solutions plus spécialisées comme **Tailor Brands** (2014, $5m) qui aide à créer son logo et sa charte de communication après avoir fourni les attributs de sa marque. L'agence de communication **McCann** de Tokyo utilise de son côté une IA comme directeur de création. Tout cela est bien vaporeux car pas très bien documenté.

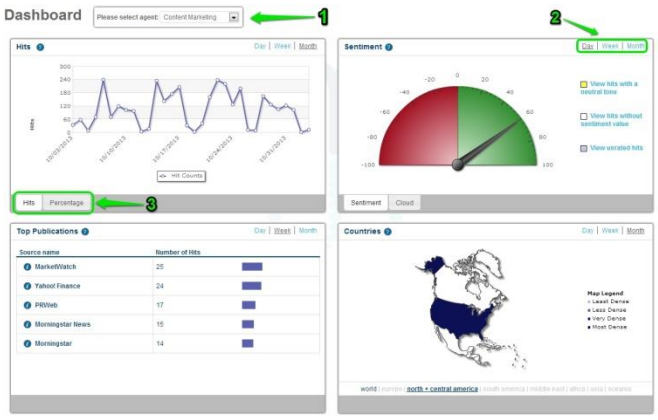


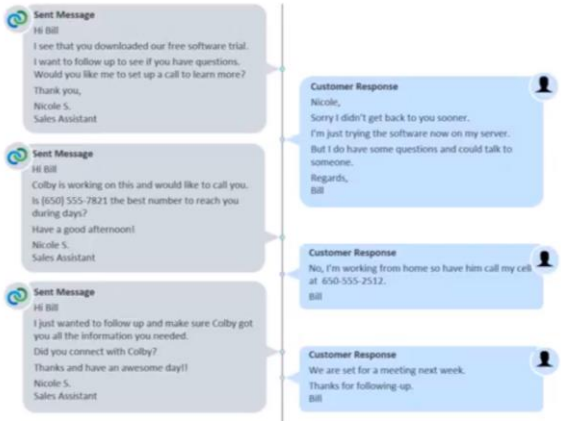
Nous avons aussi des solutions d'optimisation de sites web comme **Webpage.ly** (2015) qui est focalisé sur le référencement naturel (SEO) et fonctionnement en mode cloud. **Tilofy** (2013, $1m) est de son côté une solution solution de prévision des tendances dans le fashion et les usages, dont les méthodes et techniques ne sont pas précisées mais qui doit faire appel à des techniques d'analyse du langage (NLP).

Network Insights (2006, $77,2m) 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 2016. 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, $4M).

Meltwater (2001, $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 l'anglais **Wrapidity**, 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*).

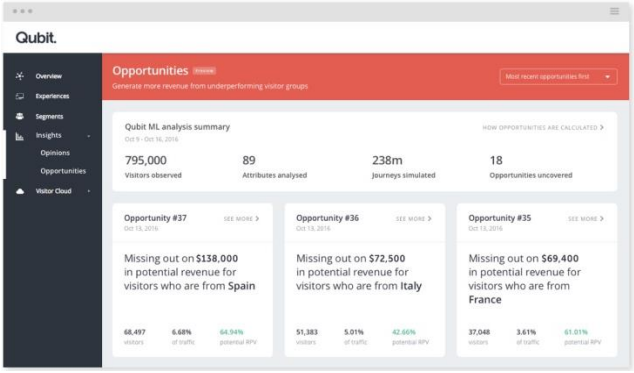


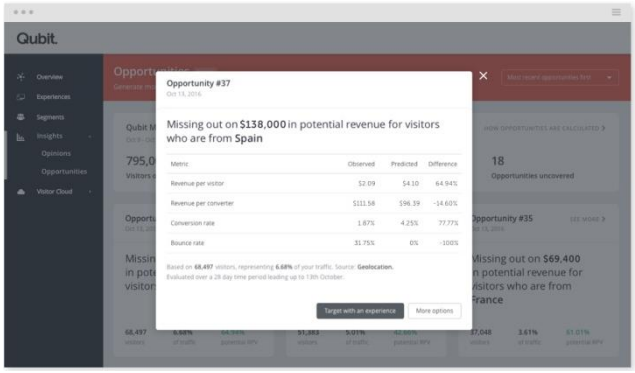


Conversica (2007, $56m) est un outil d'automatisation de la communication par mail à des prospects (*exemple ci-dessus à droite*). Comme d'habitude, les techniques d'IA ne sont pas précisées mais relèvent certainement de combinaison d'outils de NLP (traitement du langage). L'outil s'interface comme il 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.

En complément du précédent, **People.ai** (2016, $7M) 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. Pour ce qui est de visualiser leur outil, c'est une autre paire de manche : impossible à ce stade ! A part quelques vidéos de formation pour conseillers commerciaux.

L'anglais **Qubit**[133] (2010, $75m) 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.





[133] 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 ballot !

La startup française **PredicSis** (2013, $1,3M) est sur le même secteur avec ses outils de prévision du comportement des clients qui s'appuie sur des techniques de machine learning. Elle sert notamment à limiter le churn dans les entreprises proposant des services sur abonnement. Leur solution est notamment déployée chez American Express, Natixis, Crédit Agricole, Orange et EdF.

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.

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 !

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.

Nous avons tout d'abord des outils d'analyse prédictive pour identifier des talents à chasser avec **Entelo** (2011, $41M) et **Gild** (2011, $26mn). 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 startupe française **Clustree** (2013, $11,5M), lancée par Bénédicte de Raphélis Soissan, utilise aussi l'IA pour rapprocher l'offre et la demande.

Nous avons aussi des outils d'aide à la rédaction d'annonces d'emplois efficaces et d'analyse des réponses des candidats comme chez **Textio** (2014, $29,5M) (*exemple ci-dessous*). On est ici dans le domaine du traitement du langage (NLP).

textio

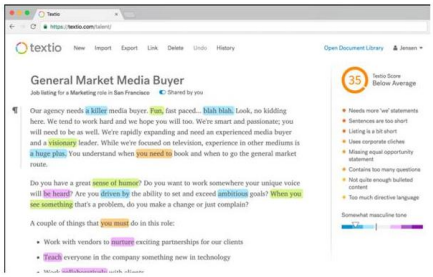
analyse automatique d'efficacité d'annonces d'emplois

\+

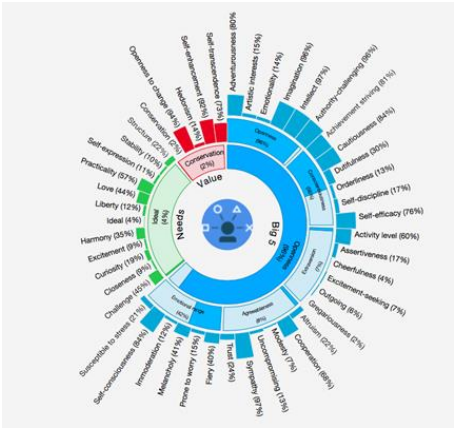
aide à la rédaction

startup US

$29,5m levés



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[134] 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.



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, $93M). Leur logiciel analyse les visages et identifie des traits de personnalité. La solution est déployée chez Unilever aux USA[135]. 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[136].

Le recrutement est aussi un autre terrain de jeu pour les chatbots comme celui que l'agence française **TheChatbotFactory** a déployé à la BNP.

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 vulnérabili-

---

[134] Cf IBM Watson Developer Cloud, Personality Insights et IBM Watson Developer Cloud, Tone Analyzer.

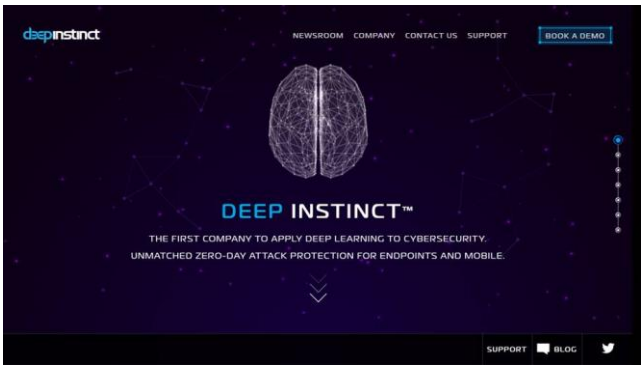
[135] Cf l'étude de cas documentée par HireVue : https://www.hirevue.com/customers/unilever-finds-top-talent-faster.

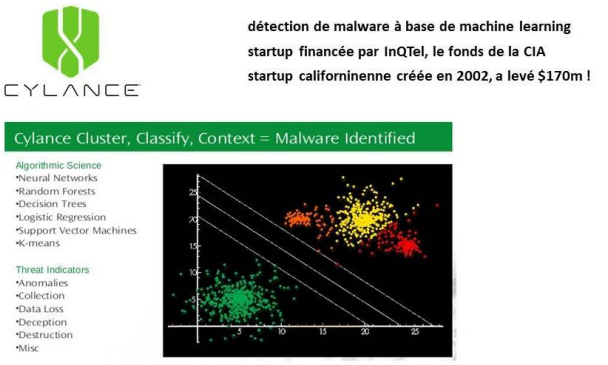
[136] Cf This AI knows whether you're gay or straight by looking at a single photo, septembre 2017.

tés diverses dans les réseaux et systèmes d'information, notamment les « non malware attacks » (remote login, attaques par scripts et macros, etc).

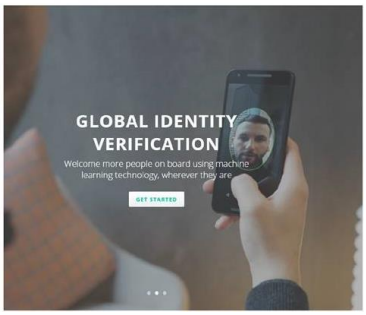
Les tentatives de phishing sont détectées par **GreatHorn** (2015, $8,83m) ou avec **Lookout** (2007, $282m) qui sécurise les mobiles avec un modèle prédictif. Les malwares sont détectés avec du machine learning par **Cylance** (2012, $177M).

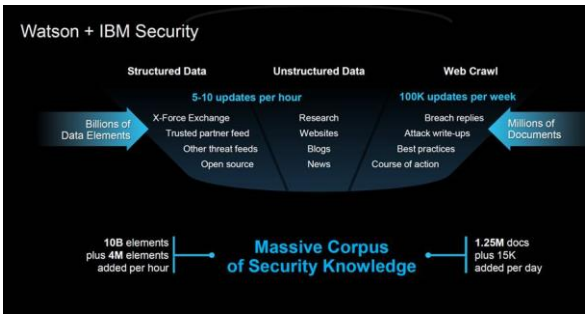
L'israélien **DeepInstinct** (2014, $32M) protège les systèmes contre les failles de sécurité récentes ("zero day threats"). Ce serait la première startup à exploiter le deep learning - avec des GPU Nvidia - tandis que la plupart utilisaient du machine learning jusqu'à présent pour faire de l'analyse multifactorielle des menaces en lieu et place de l'utilisation de bases de signatures de virus. Dans le même genre, **Recorded Future** (2009, $33M) utilise le machine learning pour détecter les menaces de sécurité en temps réel.



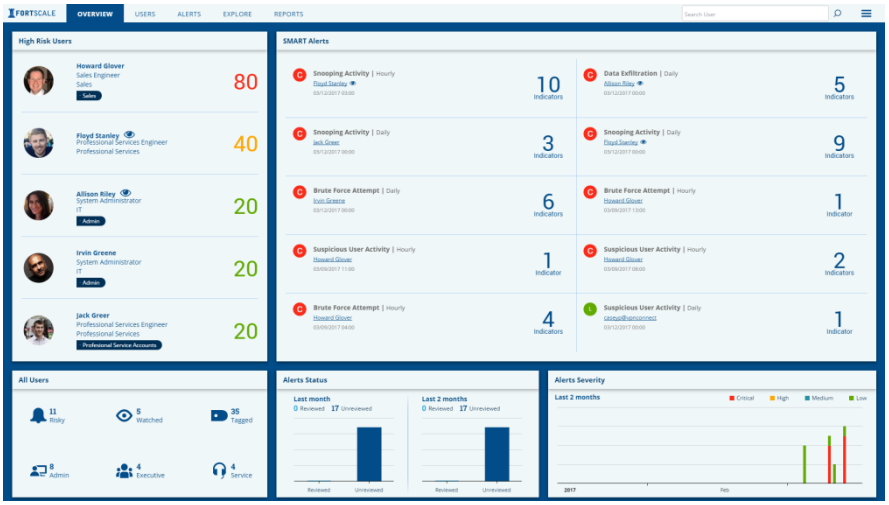


Des startups comme **Onfido** (2012, $30M) vérifient l'identité de clients de service en ligne. C'est de la détection de fraude basée sur du machine learning et du prédictif.





L'israélien **Fortscale** (2012, $32M) 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 autorisée !



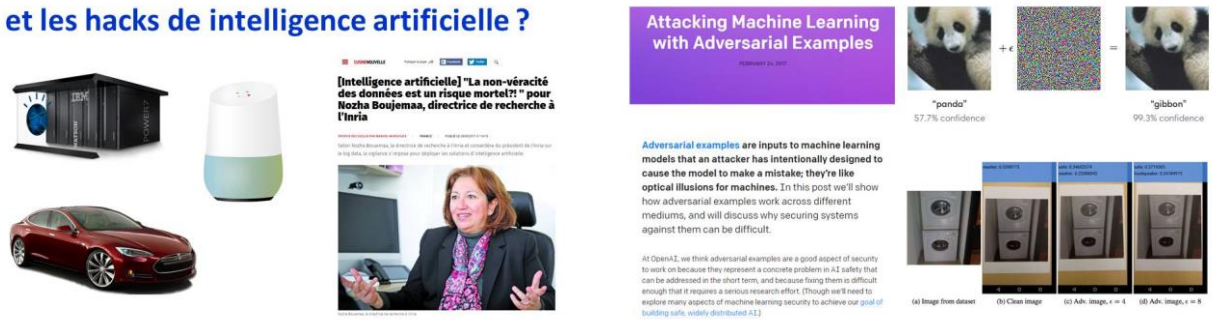
IBM qui met Watson à toutes les sauces l'a aussi décliné dans la cybersécurité. Leur Cognitive Soc (Security Operations Center) analyse toute la littérature disponible sur la cybersécurité pour aider les entreprises à détecter et circonscrire les menaces. Cignitive Soc est relié à trois outils d'IBM : BigFix Detect qui détecte les attaques dans les noeuds de réseaux, QRadar Advisor qui analyse les incidents et IBM Resilient, qui 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, $38,9M). 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.

L'israélien **Beyond Security** (1999) 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, $3,32M) pour détecter la fraude aux cartes SIM en exploitant les logs d'appels via un réseau de neurones utilisant un autoencodeur.

L'intelligence artificielle va générer de son côté de nouvelles menaces. En effet, les algoritgmes 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.

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-dessous à droite*)[137].



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.

Bref, c'est comme dans l'armement. Des mesures de défense 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 !

[137] Cf Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples, février 2016.

Applications métiers de l'IA

Dans cette grande partie, nous allons faire un tour d'horizon de l'usage de l'intelligence artificielle dans un bon nombre de marchés verticaux. Dans la pratique, ils sont tous concernés de près ou de loin. 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 ou la santé.


transports


santé


manufacturing


finance


assurance


agriculture


utilities


distribution


médias

tourisme


juridique

education


services publics


défense et renseignement

Dans ce qui suit manqueront les secteurs de l'énergie, celui de l'éducation et celui des jeux vidéos que j'aurais l'occasion de traiter dans une version ultérieure de ce document.

Ici encore, l'inventaire couvrira des projets clients et les offres de startups. 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.

Transports

Le marché des transports est vaste avec l'automobile, le transport aérien, le transport 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 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.

Mais 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.

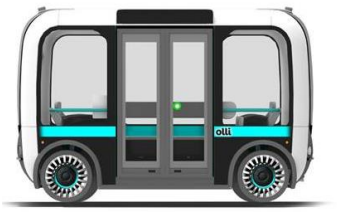
Véhicules autonomes

Les véhicules autonomes sont une catégorie de robots capable 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.

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 contraintes et difficultés que de se mouvoir dans l'espace en 3D et d'interagir avec l'environnement physique.

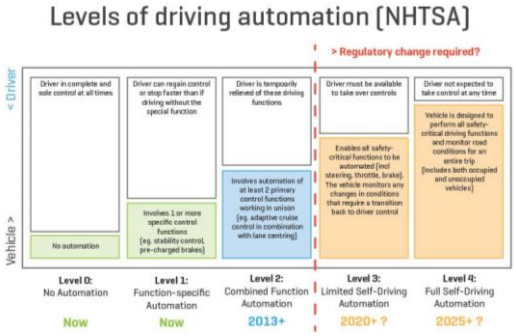
Dans une voiture, la surface de contact est relativement simple et limitée : un plan et des roues ! Plus est faible le nombre de degrés de liberté, plus l'automatisation est facile à gérer. C'est pour cela que les métros automatiques comme les lignes 1 et 14 à Paris sont déjà monnaie courante ou que les avions volent le plus souvent en pilote automatique.

Cela explique aussi pourquoi des minibus autonomes comme ceux du français **Navya** peuvent circuler dans certains lieux publics où les autres véhicules ne circulent pas. L'hétérogénéité des véhicules amène une complexité à laquelle les voitures autonomes doivent faire face.



Local Motors Olli
imprimé en 3D, dialogue avec passagers géré avec IBM Watson IoT

Navya Arma
autonome



La voiture automatique est une « réalité progressive ». Elle existe. Elle est démontrée. Si elle n'est pas encore courante, son contexte d'utilisation crédible s'agrandit d'année en année. On passera très graduellement de l'autoroute en conduite semi-autonome à la conduite en route traditionnelle puis en ville. Elle méritera alors pleinement son appellation d'auto-mobile !

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'hétérogénéite, on interdira à long terme la conduite manuelle. C'est le stade 5 de la conduite autonome.

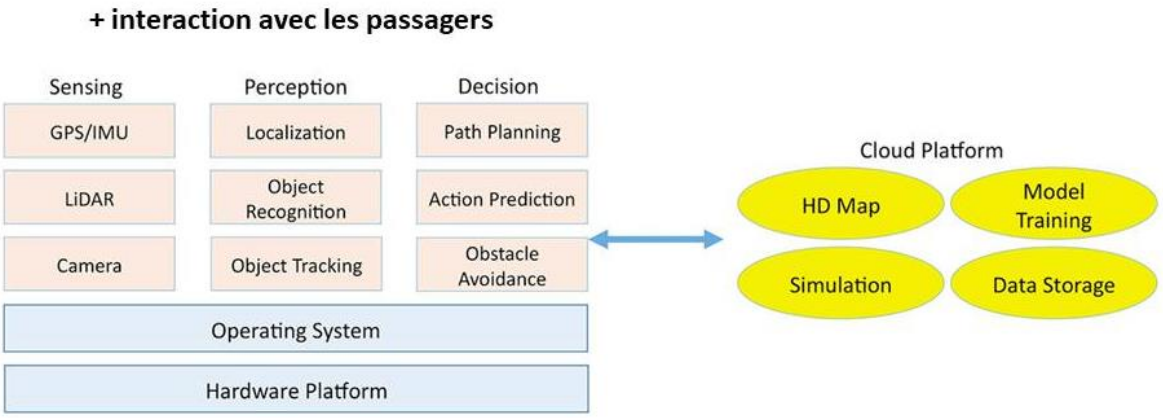
Les progrès de ces dernières années sont significatifs. Ils résultent d'efforts qui ont démarré en 2004 lors de défis lancés aux chercheurs par la DARPA. Les premiers vé-

hicules autonomes de chercheurs roulaient convenablement pendant à peine quelques dizaines de secondes.

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.

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 sont 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.



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 systèmes de vision artificielle (Mobileye, Nvidia, …), des télécommunications (la 5G jouera un rôle clé), 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 tour-

nant 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 start-tups comme Quanergy ou LeddarTech 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 l'israélien **Mobileye** qui vient d'être acquis par Intel pour $15B.

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. Elle s'améliorera sans doute avec les progrès à venir de ces GPU.

Tesla est probablement le constructeur qui a le plus de véhicules semi-autonomes en circulation avec ses Tesla S et le mode Autopilot qui est régulièrement mis à jour. Le premier accident mortel d'un véhicule ainsi équipé est intervenu mi 2016.

Un procès s'en est suivi qui a dédouané Tesla. Le conducteur n'avait pas respecté les consignes de sécurité et les alertes. 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 construc-

teur a donc fait évoluer sa configuration en multiolant les capteurs, passant notamment d'un à huit capteurs RGB. L'absence de LiDAR pourrait cependant reste un handicap.


Tesla

Tesla driver dies in first fatal crash while using autopilot mode

The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky

Danny Yadron and Dan Tynan in San Francisco

Friday 1 July 2016 00.34 BST

9557

Save for later

Joshua Brown, the first person to die in a self-driving car accident. Photograph: Facebook

juin 2017

Tesla crash report shows driver kept his hands off the wheel nearly the whole trip

Tesla S sensors

1 RGB camera

=> 8 for 360° viewing

1 radar

12 ultra-sound

+

Nvidia K1

=> Nvidia Drive PX 2

La fonction Autopilot qui a des équivalents chez d'autres constructeurs n'est pas la seule qui automatise certaines tâches de la conduite.

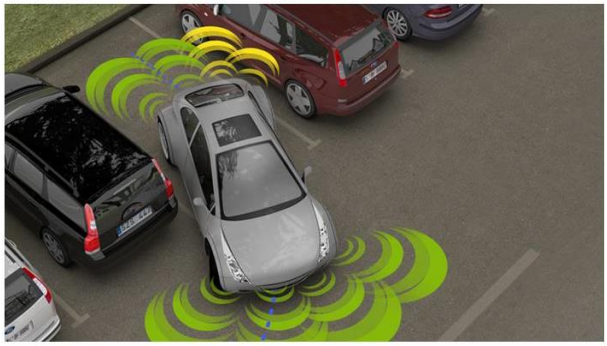
On peut aussi compter sur :

- 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.

parking automatique

disponible sur modèles haut de gamme depuis plusieurs années

exemple avec Valeo Park4U



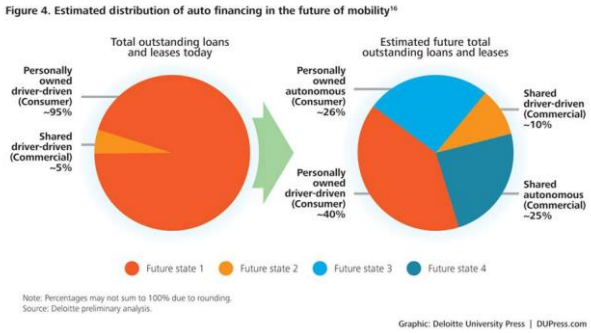
Un grand débat a cours au sujet de la gestion des dilemmes par les véhicules autonomes en cas d'accident, lorsqu'il leur faudra choisir entre la mort certaine du conducteur, de ses passagers et de personnes sur la route ou entre plusieurs personnes différentes sur la route (enfants, adultes).

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 dans la pratique[138]. 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[139].

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.

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 exemples :

- Il y aura **moins de voitures en circulation** et moins d'embouteillages dans les villes dominées par les 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**[140], avec un impact sur le marché des assurances d'un côté et aussi, sur les systèmes de santé.
- 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.

---

[138] Cf 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.

[139] 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 ! Cf A Voting-Based System for Ethical Decision Making, septembre 2017.

[140] La route fait 1,3 millions de morts par an dans le monde, concentrés en Inde et en Chine en volume, puis aux USA ! Plus que n'importe quelle guerre. Il faut y ajouter entre 20 et 50 millions de blessés. Cf http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics.

- Les **parkings** pourront être plus compacts, ceux-ci ne nécessitant pas d'être accessibles par les passagers.
- 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.

En tout cas, ils se mettent tous en branle pour se préparer à cette évolution radicale du marché. 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. **Renault Nissan** s'est aussi lancé, l'annonçant au CES 2017. **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 l'américain **nuTonomy** (2013, $19,6M) qui développe les logiciels de pilotage également utilisés par Ford.

ARGO AI

JV avec Carnegie Mellon

CEO Bryan Salesky de Google Car + anciens d'Uber

$1B investis sur cinq ans

objectif : sortir un véhicule autonome de niveau 4 en 2021

Ford a aussi investi dans Velodyne, Civil Maps et acquis SAIPS, une startup spécialisée dans la vision artificielle

Voici quelques autres startups du secteur, qu'il est très difficile de départager. Ils utilisent généralement les mêmes bases technologiques :

- **Optimus Ride** (2015, $5,25M) une spinoff du Massachusetts Institute of Technology qui développe la partie logiciel de véhicules autonomes de niveau 4.
- **Netradyne** (2015, $16M) est un spécialiste de deep learning appliqué la la vision des véhicules autonomes avec leur plateforme logicielle Driveri.
- **Drive.ai** (2015, $62M) a été créé par des anciens de Stanford qui veulent aussi proposer une plateforme de conduite autonome à base de deep learning.
- **Comma.ai** (2015, $3M) ambitionne de proposer une sorte de SDK permettant de rendre autonomes des véhicules existants.

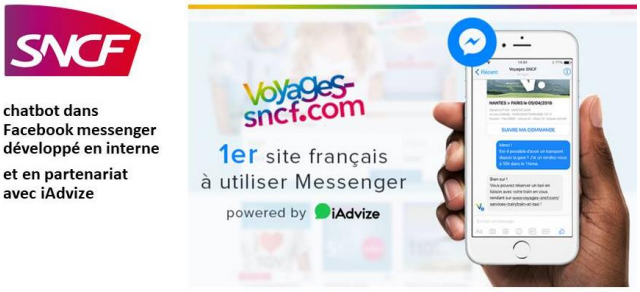
Autres usages

Dans les transports, l'IA a d'autres usages que la conduite autonome.

Elle peut servir à optimiser les trajets, notamment 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. Enfin, l'analyse du bruit du moteur permet de faire de la maintenance préventive. C'est une fonction aussi proposée par la startup **Otosense** (2014), qui est basée à Cambridge (USA) et créée par le français Sebastien Christian.

La maintenance prédictive est aussi un sujet d'intérêt pour la **SNCF**, avec **Quantmetry**, Global Sensing Technologies et IBM Watson. Cette même SNCF utilise des chatbots dans les résaux sociaux.



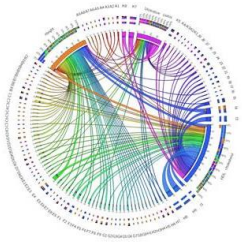
Enfin, dans le transport aérien, la robotique joue un rôle émergent avec par exemple, le **BagBot** qui remplit automatiquement les containers de valises dans quelques aéroports européens depuis 2014 et le **Skywash** qui lave les avions de toute taille depuis 1997, notamment à Frankfort en Allemagne.

DigiBot BagBot (2014) Schiphol

Skywash (1997) Putzmeister

Santé

C'est le marché vertical le plus attirant pour les startups de l'IA avec celui de la finance et du commerce. L'IA est notamment utilisée dans la génomique et dans l'aide au diagnostic dans la lignée de la solution en oncologie que nous avons vue au sujet d'IBM Watson. Ce qui suit n'est probablement qu'une liste très partielle des startups de ce secteur d'activité prometteur.



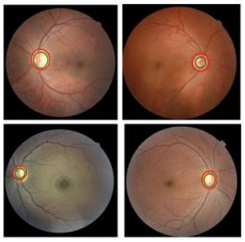
biotechs

criblage de molécules

simulation du vivant

séquençage

covariance phénotype/génotype



diagnostics

imagerie médicale

détection de pathologies diverses

ECG, EEG

systèmes de santé

médecine prédictive

prévention des erreurs

réduction des risques

Nous allons segmenter les usages de l'IA en trois grandes catégories : les biotechs, les outils de diagnostics et enfin, tout ce qui se rapporte aux systèmes de santé.

grande vague de startups dans l'IA et la santé !

Biotechs

L'univers de la pharmacie et des biotechs est un très gros consommateur 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.

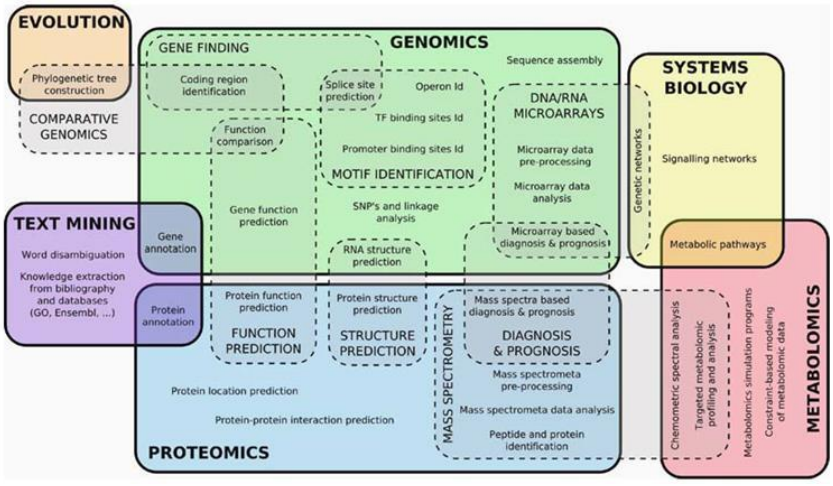
L'IA peut aider à comprendre la structure des gênes et de leur expression, l'épissage des gênes (comment les différentes parties d'un gêne s'assemblent), le repliement des protéines sur elles-mêmes après leur production dans les ribosomes des cellules ou la

détermination des paramètres qui favorisent ou pas l'expression des gênes, notamment ceux qui déclenchent des cancers.

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[141].

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[142]. 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.

Bien qu'utilisés essentiellement en imagerie médicale, les réseaux de neurones convolutionnels (ConvNets ou CNN) 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).



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

Nous ne traiterons pas ici du sujet dans ses recoins mais plutôt l'illustrer par quelques startups actives dans le domaine comme presque partout ailleurs dans ce document.

Innoplexus (2011) est une startup indo-allemande qui propose son moteur de recherche iPlexus d'informations médicales et exploite 27 millions de publications, 365 000 rapports de tests cliniques et un million de thèses. Le tout exploite du machine learning et du traitement du langage.

[141] Source : « Deep Learning in Bioinformatics » des coréens Seonwoo Min, Byunghan Lee et Sungroh Yoon (2016).

[142] Cf Scientists develop machine-learning method to predict the behavior of molecules, octobre 2017.

Iktos

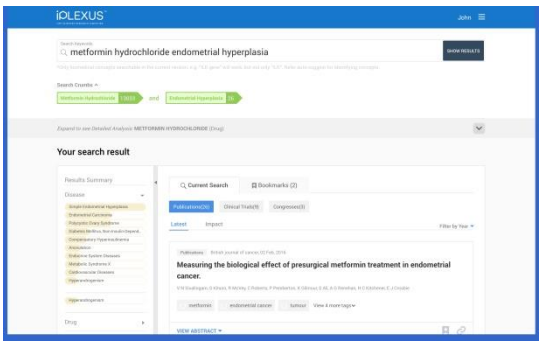
startup française
identification de molécules thérapeutiques
langage mathématique de description de molécules
deep learning pour criblage
nombreux concurrents...

optibrium

ChemAxon

CERTARA

Mind the Byte



IKTOS (2016) est une startup française toute récente qui 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. Ils ne sont pas seuls sur ce marché qui comprend d'autres startups telles que Certara, ChemAxon, Mind the Byte, Optibrium et Tripos.

Atomwise (2012, $6,35m) 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.

La startup française **Owkin** (2016, $2,1M) 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.

Insilico Medicine (2014, $14M) 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).

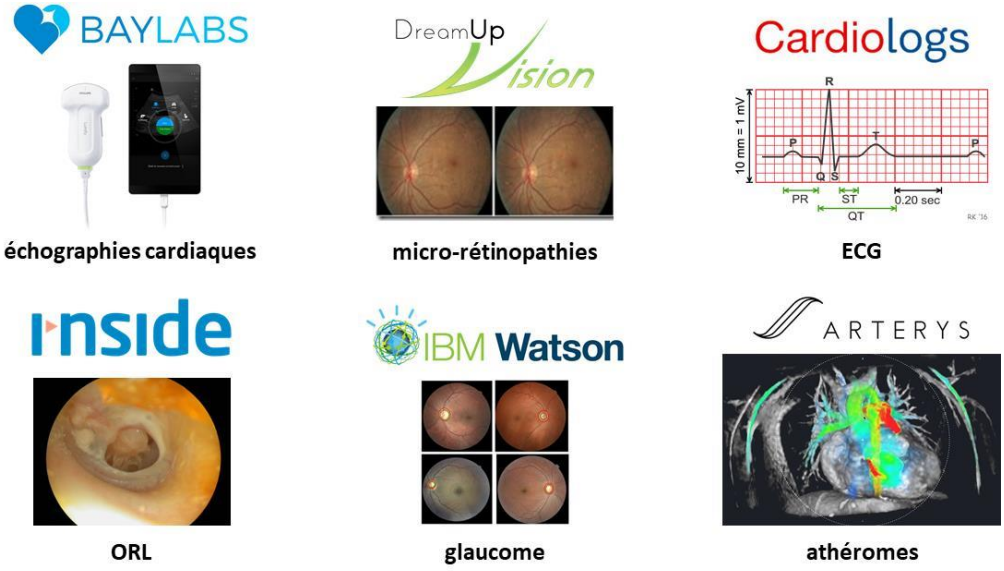
Aide au diagnostic

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 omniprésents dans l'interprétation d'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.

Ces 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'entrainement 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. Elles peuvent aussi comparer avec précision des images dans le temps en tenant compte des variations de prises de vues.

Pour être entrainés, ces solutions doivent évidemment exploiter des bases d'imagerie déjà tagées 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.

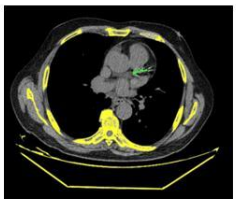


Quasiment tous les pans de l'imagerie médicale sont touchés par le deep learning pour la détection de pathologies :

- **Microrétinopathie** par analyse du fonds de l'œil, surtout pour les diabétiques avec le français **DreamUpVision**.
- **Glaucome** avec **IBM Watson** en Australie, une des principales causes de cécité dans le monde qui est détectée trop tardivement dans la moitié des cas. L'entrainement est basé sur l'exploitation de 88 000 fonds de rétines[143].
- **Pathologies de l'oreille** comme les oreillons et éclatements de tympans par analyse du tympan avec un otoscope avec le français **i-nside** qui a entrainé sa solution de deep learning avec 250 000 images.
- **Ostéoporose** par analyse de radios avec l'Israélien **Zebra Medical Vision** (2014, $20M) qui détecte aussi les compressions de vertèbres, la stéatose hépatique (foie gras) et les hémoragies cérébrales.

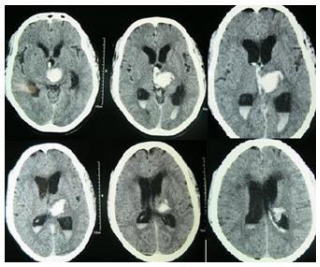
[143] Source : https://www.ibm.com/blogs/research/2017/02/watsons-detective-work-could-help-stop-the-silent-thief-of-sight/?glaucoma

ostéoporose



stéatose hépatique

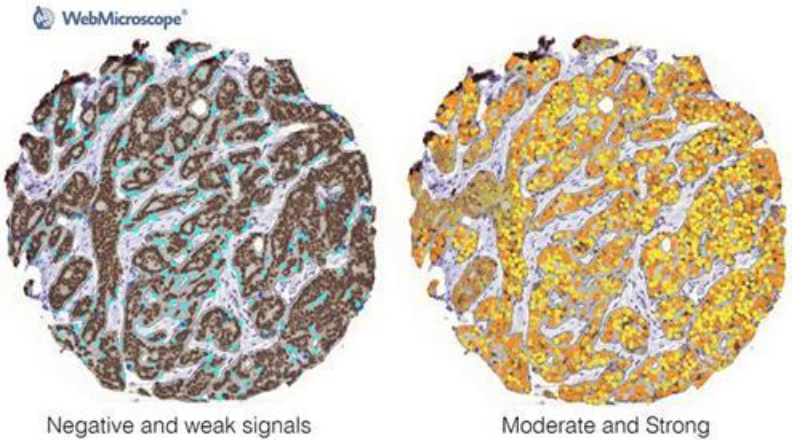
compression de vertèbre



hémoragies cérébrales

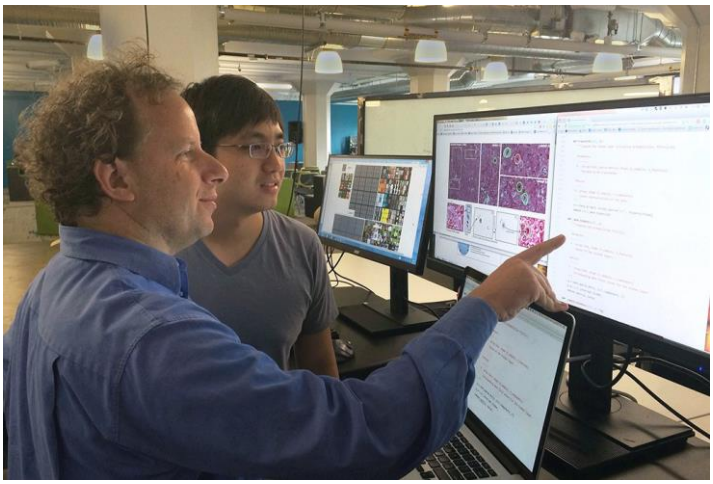
- **Cancers du poumon** et **fractures osseuses** par analyse de radios avec **Enlitic** (2014, $15M) qui est un généraliste de l'exploitation de radios. C'est aussi le champ d'action de **Riverain Technologies** qui est focalisé sur la cage thoracique.
- **Pathologies cardiaques** par analyse d'échographies avec **Baylabs** (2013) et chez **Arterys** (2011, $13,72M) ainsi que chez l'israélien **DIACardio** (2009, $2M)
- **Pathologies du sein** avec **Volpara Solution** (2009, $5,5M + IPO) qui réalise des analyses densitométriques précises et aussi **QVCad** (2006, $4,75M) ainsi qu'avec le français **Therapixel** (2013, $600K). Ces systèmes peuvent notamment être entrainés 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[144].
- **Pathologies du cerveau** comme avec 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. Et aussi avec **MedyMatch** (2013, $2M) qui détecte les hémoragies cérébrales et exploite les briques de reconnaissance d'image d'IBM Watson.
- **Analyse de cellules cancéreuses dans des biopsies** avec le Finlandais **WebMicroscope** (2013, $1,7M) 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 aussi un résultat quantitatif par comptage de cellules (*exemple ci-dessous*).

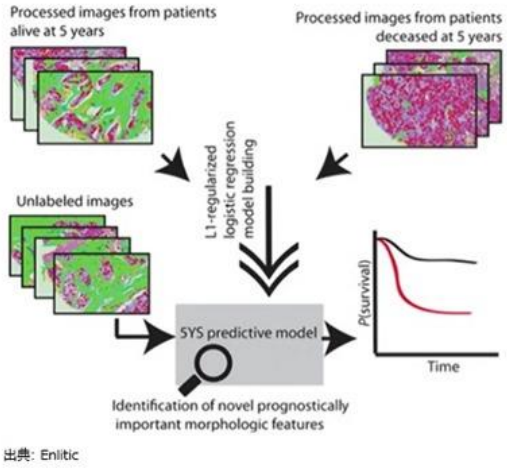
[144] Cf https://www.synapse.org/#!Synapse:syn4224222/wiki/401743.



De nombreuses startups se positionnent comme des généralistes couvrant plusieurs pathologies :

- **Enlitic** (2014, $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[145].

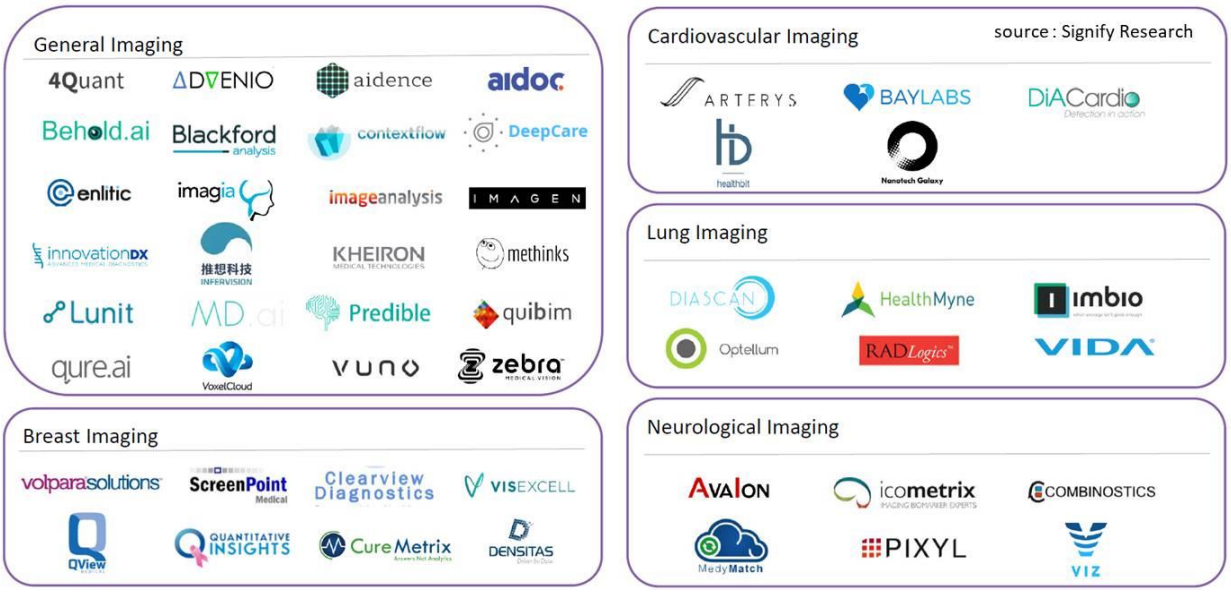




- **Behold.ai** (2015, $20K) a développé une solution d'analyse d'imagerie médicale pour aide les radiologues à faire leur diagnostic. Cela s'appuie sur du machine learning. Le système compare les images de radiologie avec et sans pathologies pour détecter les zones à problèmes, comme les nodules et autres formes de lésions. **HealthMyne** (2013, $11,5M) propose aussi un logiciel généraliste analyse de radios qui produit des rapports quantitatifs sur certaines observations.
- **VoxelCloud** (2015, $13,5M) qui couvre le cancer du poumon, la rétinopathie diabétique, les maladies coronariennes et du foie.

[145] Cf 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.

- **Lunit** (2013, $5.5M), une startup coréenne qui propose une solution logicielle de deep learning générique d'interprétation d'imagerie médicale, notamment de radios et qui semble commercialisée en OEM.
- **Butterfly Network** (2011, $100M) est en train de créer 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 à 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[146].

L'imagerie médicale n'est pas la seule source de diagnostics médicaux. Il faut ajouter l'analyse d'ECG (électrocardiogramme) et d'EEG (électroencéphalogrammes), les tests en laboratoires (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.

Cardiologs (2014, $1.2M) 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 utili-

[146] Cf 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.

sent comme nombre de startups de l'IA les ressources de Google Tensorflow. Cela permet de fournir une réponse en quasi-temps réel. Ils ont entrainé 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.

Dans la pratique, l'analyse d'ECG passe par l'utilisation des plusieurs méthodes et différents réseaux de neurones (récurrents, convolutionnels, autres)[147]. L'efficacité des méthodes dépend de la pathologie à détecter. Par exemple, la fibrillation attriale est mieux détectée avec un réseau convolutif.

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].

HealthReveal (2015, $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.

FORWARD

startup de San Francisco
centre de soins new wave
doté de tous les capteurs du jour
outils de machine learning pour l'aide au diagnostic

HealthReveal

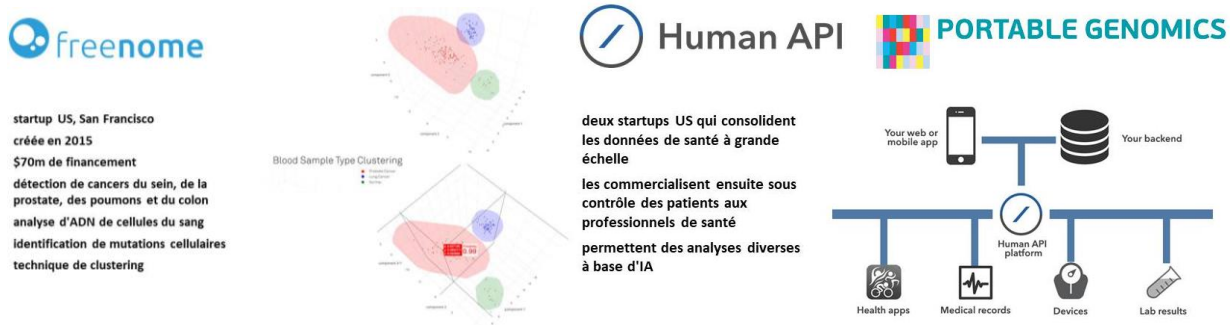
startup US créée en 2015
financement de $10m
stealth mode
prévention de maladies chroniques
principalement diabète et cardio-vasculaire
pour tiers payants et patients
technologie non indiquée, probablement mix de machine learning et de deep learning


HealthReveal Announces $10.8 Million Series A Funding

Investment supports convergence of advanced analytics and the clinical Internet of things to preempt the consequences of chronic disease

[147] Voir Machine Learning in Electrocardiogram Diagnosis, 2009.

Forward (2016) est une étonnante startup américaine qui veut inventer le cabinet médical. Son premier site à San Francisco est équipé de tous les capteurs, outils d'analyses de laboratoires, ADN compris, et systèmes d'imagerie médicale pour faire un bilan de complet à 360°[148]. Ce n'est pas une clinique pour autant.



Freenome (2015, $70M) est une société de San Francisco qui produit des analyses de biopsies liquides – essentiellement du sang - permettant la détection de cancers émergents (vidéo). Cela repose probablement sur de l'imagerie de cellules sanguines après exposition à des marqueurs chimiques et génomiques.

Portable Genomics (2011) 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, indispensable aussi bien pour les praticiens que 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, $3,7m) 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

[148] 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/.

gênes et de leurs effets sur leur épissage[149]. 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, $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).

Sophia Genetics (2011, $58,75M) est une startup suisse qui 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, $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 autotraitement de certaines pathologies par les patients.

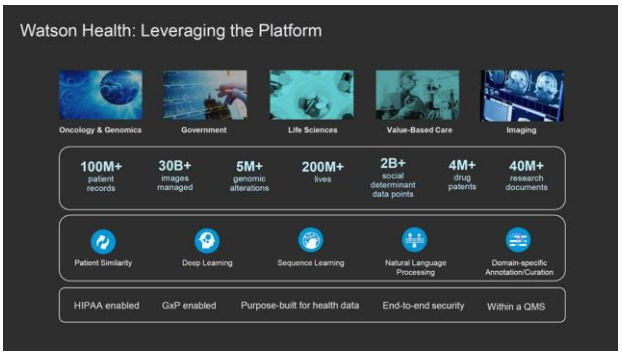
Lumiata (2013, $20M) est dans la même lignée un système d'analyse de situation de patient permettant d'accélérer les diagnostics, notamment en milieu hospitalier.

MedWhat (2010, $560K) 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 !

Nous allons terminer ici avec **IBM Watson** qui est décliné sur un grand nombre de cas d'usages et notamment sur l'aide au diagnostic du cancer en partenariat avec de nombreuses cliniques américaines ainsi qu'à l'étranger, et en exploitant les données phénotypiques (présence de la maladie) et génotypiques du patient et les bases de connaissance du secteur composées de millions de documents (recherches, études cliniques, etc). Ces outils destinés aux cancérologues ont été largement survendus par le marketing d'IBM et ne sont pas encore véritablement éprouvés à grande échelle.

[149] Voir **The human splicing code reveals new insights into the genetic determinants of disease** qui explique les fondements scientifiques de leur procédé.

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.



Elle a ensuite été déployée 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. Elle est fournie sous forme de service en cloud, avec un abonnement dont le prix n'a pas été rendu public par IBM.

There Are Barriers to Adopting IBM's "Cadillac" Solution. Our checks suggest that IBM's Watson platform remains one of the most complete off-the-shelf platforms available on the marketplace. However, many new engagements require significant consulting work to gather and curate data. Our checks suggest that Watson is a finicky eater when it comes to data enterprises can feed it – in other words, IBM has very exacting standards for data preparation. The halt of and cost overruns in the MD Anderson engagement with Watson epitomize our concerns here.

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center

projet IBM Watson lancé en 2013 et abandonné en 2017, d'abord sur les leucémies puis sur le cancer du poumon

solution jamais été utilisée sur des patients

le projet a coûté $63m,versés à IBM et PricewaterhouseCoopers

"The problem isn't too much data for humans–it's too little"

La solution analyse les dossiers de patients atteints de tumeurs cancéreuses, y compris le séquençage d'ADN des tumeurs[150], aide au diagnostic, détermine des traitements possibles et évalue leur efficacité relative.

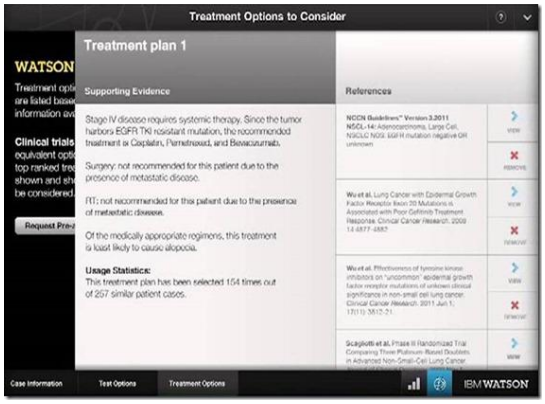
Elle aide notamment à optimiser l'usage de la chirurgie, de la radiothérapie et de la chimiothérapie. Les cancers sont des pathologies idéales pour Watson car elles sont plurifactorielles. Mais ce n'est pas (encore) de la médecine préventive.

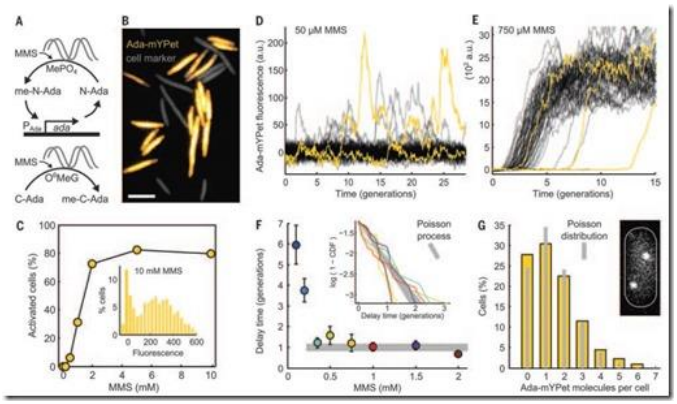
Les études scientifiques publiées sont très nombreuses et toujours fournies avec des résultats statistiques sur des cohortes de patients. Il faut les croiser avec des logiques statistiques et cognitives complexes pour en tirer des conclusions. On connait par exemple le lien entre les mutations des gênes BCRA1 et BCRA2 et les cancers du sein.

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

[150] 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.

données relationnelles, etc). Il semble que cette partie de la solution ait été développée en partenariat avec **Cleveland Clinic**.





Watson utilise des sources d'informations variées pour faire son diagnostic, et il 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 relativement facile à analyser, autant les illustrations qui ne sont pas fournies sous format structurées comme dans l'exemple ci-dessous, ne doivent pas être 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 Watson sait tenir compte de la forte proportion de publications scientifiques qui est entachée de fraudes ou exagérations[151].

Dans les démonstrations, la solution 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.

---

[151] 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.

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, recherche 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 faudra 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. Dans la pratique, les annonces évoquées ci-dessus ne sont pas vraiment éprouvées ni déployées à grande échelle[152].

Dans les applications santé de Watson, on peut aussi citer l'application de **GenieMD** 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.

En 2014, le **Baylor College of Medicine** a créé son application KnIT (Knowledge Integration Toolkit) à base de Watson pour identifier des therapies 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 !

Enfin, 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.

---

[152] Cf 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.

Systèmes de santé

Les solutions de systèmes de santé couvrent des besoins divers : l'observance des traitements, les robots chirurgicaux[153], les robots pour s'occuper des personnes âgées, l'évitement d'erreurs de prises de médicaments, 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.

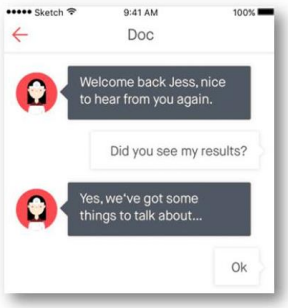
En voici quelques exemples, toujours pris dans l'univers florissant des startups :

Cognitive Scale (2013, $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. Le français **Diabeloop** propose une solution complète intégrant un capteur de glycémie en continu et une pompe à insuline et un logiciel de suivi exploitant lui aussi des briques d'IA.

MedAware (2012, $10,3M) 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 exploite notamment des bases de données médicales d'historiques de patients.

Hindsait (2013) propose 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 !

Doc.ai (2016) propose un chatbot de dialogue avec un médecin virtuel, dédié à l'analyse de résultats de laboratoires, ce qui n'empêche pas, ensuite, d'aller voir un médecin et surtout un spécialiste.



OrCam Technologies (2010, $56M) 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

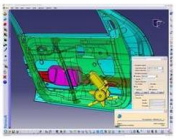
[153] La plupart des robots chirurgicaus sont télécommandés comme les Da Vinci de la startup américaine **Intuitive surgical** qui sont spécialisés dans les opérations de l'abdomen et sont déployés depuis plusieurs années.

to speech. La caméra miniature se positionne sur des lunettes traditionnelles et comprend un écouteur.

La startup française **Panda Guide** (2015) est positionnée de son côté avec 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 entrainé sur serveur, capable de rcconnaitre un millier d'objets de la vie courante.

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.



conception et simulation
CAO, PLM, VR
mise en situation

fabrication
gestion stocks et entrepots
robots de fabrication et assemblage
contrôle qualité par imagerie

exploitation
objets connectés
metering
maintenance prédictive
gestion d'assets

AUTODESK.

"design génératif" avec Dreamcatcher dans Netfabb 2018

à base de Machine Learning pour proposer diverses formes

Nous avons déjà rapidement évoqué les robots de production dans la rubrique sur les robots.

Autodesk a récemment développé autour de sa solution Dreamcatcher une fonctionnalité à base de machine learning qui permet de choisir différents designs en phase de conception d'objet. Elle exploite une base de formes et d'objets qui est automatiquement adaptée à l'objet en phase de conception.

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 des ascenseurs fait appel au machine learning chez **Kone** et **Schindler**.

L'opérateur américain d'éolienne **Invenergy** exploite une solution logicielle issue de l'éditeur de logiciels **SparkCognition** (2013, $39M) avec machine learning et deep learning pour déterminer les périodes de maintenance préventive des éoliennes en avance de phase, exploitant quatre années de données sur 100 éoliennes et 27 variables de fonctionnement.

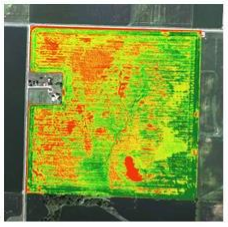
Le fabricant japonais de composants **Rohm** propose des composants d'analyse du bruit de moteurs exploitant des réseaux de neurones. Les alertes sont remontées en central via des réseaux de télécommunication bas débit.

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.

L'industrie est influencée par les applications grand public, notamment dans l'univers des objets connectés. Citons par exemple le français **Ween** et son système de contrôle du confort qui s'appuie sur une IA détectant le comportement des habitants (notamment leurs trajets) pour optimiser leur confort thermique. Ici, à base de machine learning. Et puis **Modiface** qui fait de la recommandation cosmétique en fonction de l'analyse du visage, ici avec des réseaux de neurones convolutionnels.

Agriculture

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.



télédétection

analyse imagerie drones et satellites

prévisions récoltes

agriculture de précision

exploitation capteurs

planification

optimisation des ressources

robotisation

binage, semis, traitements, récoltes, packaging

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 ou 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 la qualité prévisible des récoltes, à l'échelle de son exploitation aussi qu'à l'échelle globale, ce qui permet d'anticiper les cours de vente de sa production voire de les optimiser.

La télédétection s'appuie sur de satellites, des avions mais aussi sur des drones. La startup suisse **Gamaya** (2015, $4M) propose une caméra multispectrale pour drones, scannant une quarantaine de fréquences électromagnétiques dans le visible et non visible. Le tout pour produire des cartographies précises des champs et identifier leurs parties qui seraient atteintes de pathologies ou manquant d'irrigation.

Agriculture de précision

Bowery Farming (2015, $31M) est une startup de New York dont le système d'exploitation BoweryOS 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) est une startup allemande dont l'application mobile Plantix exploite de simples photos de végétaux prises par des smartphones pour identifier les maladies ou parasites qui les affectents. L'application fournit des recommandations. Celle-ci est gratuite et la startup espère exploiter les données récoltées, un modèle économique toujours difficile à mettre en place.

Benson Hill BioSystems (2012, $34,5M) est une autre startup américaine qui 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.

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[154], 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 technologie** 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). Une startup suisse **EcoRobotix** (2011, $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 et n'a pas été industrialisé.

[154] 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.

Le robot de surveillance et de désherbage Thorvald 2 du norvégien **Saga Robotics** permet d'éviter de faire appel à des pesticides (vidéo, *ci-dessus 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-dessus à gauche*).

CNH Industrial a créé un concept de tracteur autonome pour 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**, 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** 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*)[155].

[155] Cette intéressante présentation décrit différentes méthodes de récolte de pommes : https://agrifoodroboticsworkshop.files.wordpress.com/2015/11/zhang2015iros_afr.pdf.

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.

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).

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) a l'air opérationnel, avec cueillette robotisée mais intervention humaine pour remplir les barquettes (*ci-dessous à gauche*).

L'américain **Blue River Technologies** (2011, $30,35m) 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** 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 apparaitre 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.

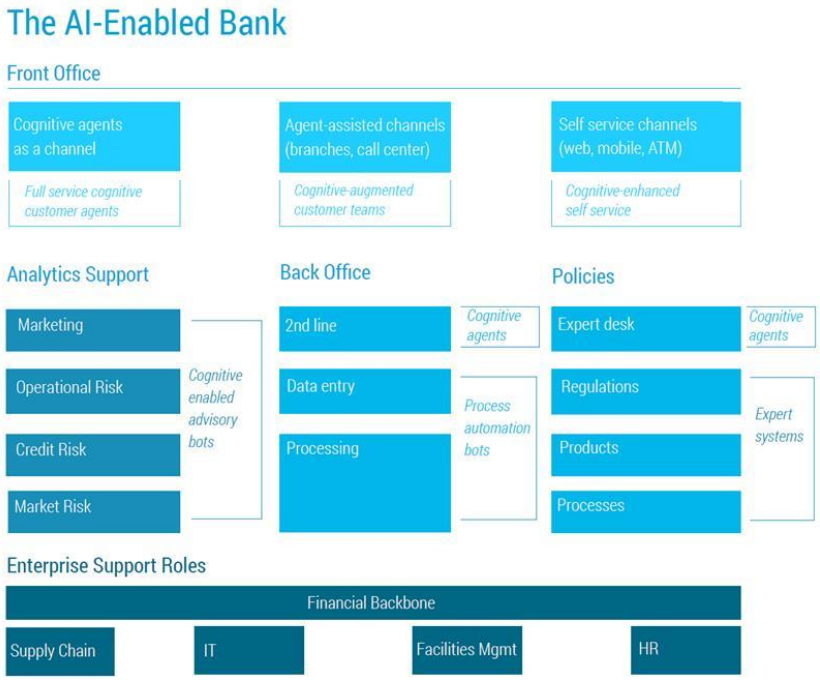
Après ces quelques recherches de robots agricoles, je me rends compte finalement que l'on en est à peu près au même stade que pour les robots humanoïdes : les démonstrations et effets d'annonce sont nombreux, mais les réalisations concrètes opérationnelles le sont bien moins. Cela ne veut évidemment pas dire que cela ne marchera jamais mais que la mise au point de ces robots agricoles dans des conditions économiques satisfaisantes est encore un long chemin semé d'embuches.

Finance

La finance est un autre de ces marchés verticaux qui croule sous les données et à ne pas savoir qu'en faire ! C'est donc un terrain très favorable à l'usage de techniques de

machine learning, dans tous les métiers de la banque et de la finance[156], du front office au back office en passant par les fonctions analytiques, 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, 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.

Ceci étant dit, 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 des clients. La banque directe se développe lentement, surtout en France. Les Fintechs ambitionnent de disrupter le marché à tous les étages (mobilité, Bitcoins, crowdfunding, etc) mais sont encore trop focalisées sur les moyens de paiement.

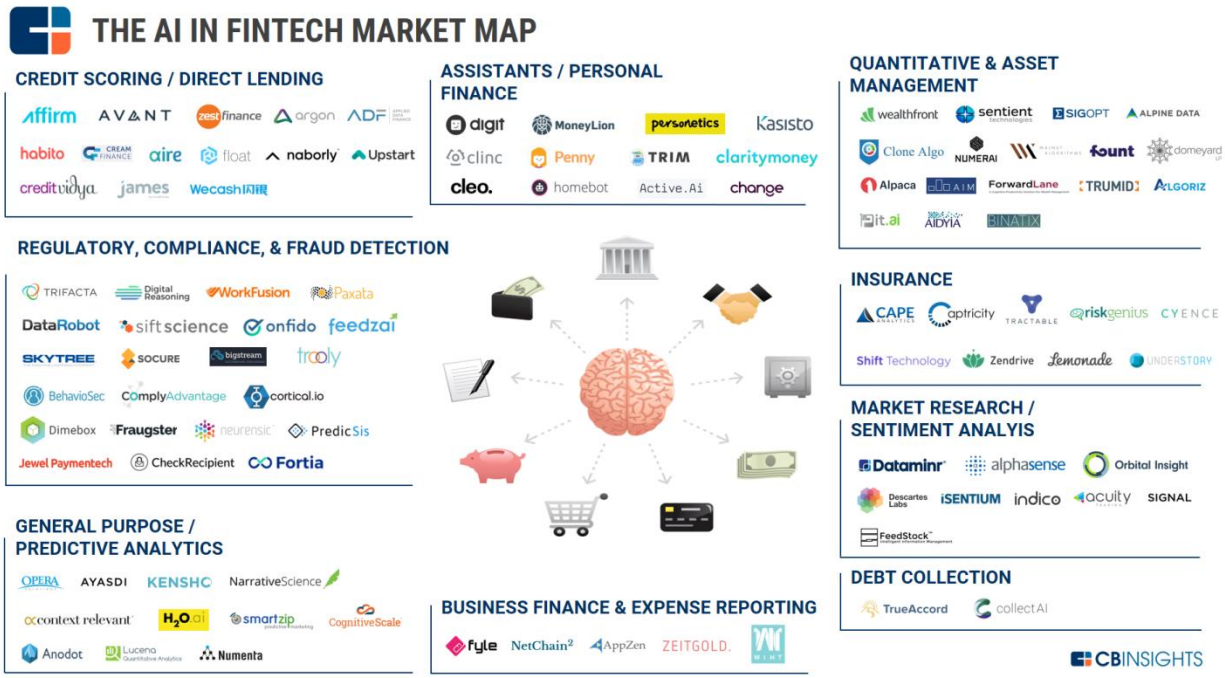
En d'autres termes, elles gèrent plus les petits flux que les gros flux de transactions et les stocks d'investissements. L'IA pourrait-elle accélérer la mutation du secteur ? Est-elle un facteur qui peut faire évoluer la notion de confiance, critique dans le secteur, d'un type d'acteur à un autre [157]? Pour l'instant, rien ne semble l'indiquer.

Les startups ont aussi un obstacle à franchir : le machine learning qu'il soit supervisé ou pas, ne peut pas produire de résultats probants sans un gros volume de données.

---

[156] Le schéma vient de Tomorrow's AI-Enabled Banking de IPSoft.

[157] 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. Cf le baromètre mondial de la confiance 2017 d'Edelman.



Chatbots

Ici 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. Le marché 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 clients.

Les startups de la première catégorie sont très nombreuses avec les anglais **Cleo** (2015, $3M), **Plum** (2016, $500K) et **Chip** (2016), et puis **Abe** (2016), **Digit** (2013, $36M), **Trim** (2015, $2,2M), **Penny** (2015, $1,2M) et **Dyme.co** (2014, $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**[158] et son chatbot MyKai (2013, $11M) qui est notamment utilisé par MasterCard, *ci-dessous à gauche,* **Finn.ai** (2014) et puis bien évidemment **IBM Watson** qui 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.

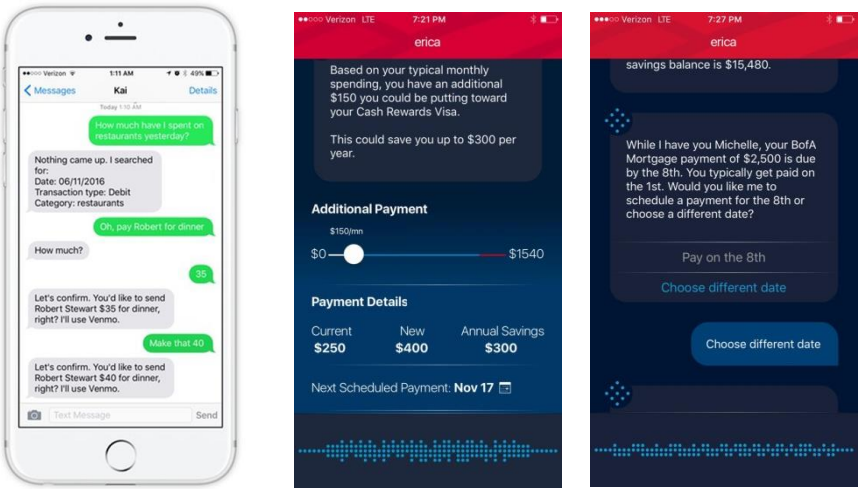
Aux USA, des chatbots ont été lancés par **Bank of America** avec Erica qui est plutôt structuré comme un système de recommandation et est aussi commandable par la voix (*ci-dessous à droite*), et **American Express** sur Facebook Messenger depuis fin 2016 et **Wells Fargo** depuis mi 2017, au Royaume Uni, par **Barclays** et son

[158] Avec plus ou moins de bonheur, voir : https://www.wired.com/2016/06/new-banking-ai-now-chatbots/.

Launchpad qui permet d'exécuter des tâches de son application mobile classique en mode dialogue, sans que cela soit d'ailleurs plus efficace et chez **Santander** en Espagne[159] ou **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**.

La banque singapourienne **OCBC** a créé un chatbot dénommé Emma (tous les prénoms féminins y passent…) spécialisé dans l'attribution de prêts pour la rénovation de logements[160].

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.



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. Dans la pratique, ils servent surtout à réduire les coûts du retail banking[161].

Optimisation d'investissements

La seconde catégorie de startups qui s'est engouffrée dans l'exploitation de l'IA à des fins financières est celle qui couvre toutes les solutions d'optimisation de gestion des investissements, surtout boursiers. La majorité des solutions sont b2c et quelques unes sont b2b.

---

[159] Voici la source de nombre de ces différents exemples : « Artificial Intelligence in Digital Bankin » de MAPA, novembre 2016.

[160] Source : OCBC bank launches first artificial intelligence powerd home & renovation loan specialist, avril 2017.

[161] Cf Bots Aren't Ready To Be Bankers de Forrester, août 2016.

L'optimisation d'investissements boursiers est proposée par **DataFox** (2013, $6,8m), **AdvisorEngine**, anciennement Vanare (2014, $26M) avec son logiciel en cloud de gestion d'investissements personnels et **WealthArc** (2015, $1,5M) et sa solution destinée aux conseils en gestion de patrimoine.

wealthfront

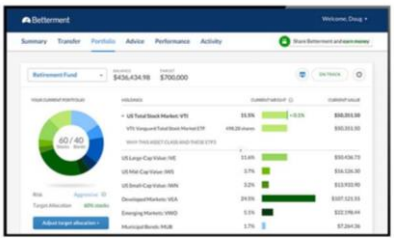
gestion financière en ligne

exploitation de machine learning pour optimiser les investissements par classe

startup US créée en 2011

$129m levés

Betterment



gestion financière en ligne

exploitation de machine learning pour automatiser la gestion de placements

startup US créée en 2008

$205m levés

La gestion financière en ligne est proposée par **Wealthfront** (2008, $129M) et **Betterment** (2008, $273M), deux startups extrêmement bien financées et qui ne ciblent visiblement que le marché US ! Et puis aussi **Pefin** (2015) qui gère également la finance personnelle des foyers de tous niveaux de patrimoines.

Pefin

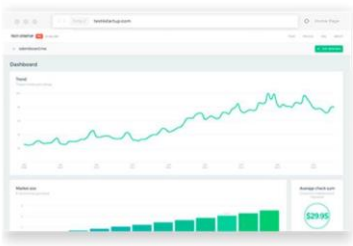
gestion financière en ligne pour les familles, entièrement gérée par logiciel et l'IA

IA à base de réseaux de neurones

startup basée à New York

On trouve même des solutions pour identifier des startups dans lesquelles investir avec **Mattermark** (2012, $17,2m) et le russe **TalentBoard** et sa solution Test4startups (ou T4S). Dans la pratique, Matermark est devenu 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.

T4S



évalue un business plan de startup

compare les hypothèses aux concurrents et à la taille des marchés visés

scanne les médias sociaux

exploite le Machine Learning d'Amazon, Microsoft et Google

destiné aux investisseurs

startup russe sous le nom Test4startups

DEEP KNOWLEDGE LIFE SCIENCES

AGING ANALYTICS AGENCY
Invest for life

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

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 en était un de ses employés. Aussi sophistiquée soit-elle, une solution d'IA reste un outil d'aide à la décision. Elle fournit des informations qui servent d'aide à la décision, comme dans n'importe quel autre processus de décision qui s'appuie sur la rationnalité de données.

La startup française **Khresterion** (2014) 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'ontologie 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.

LendUp propose des prêts aux particuliers en optimisant leurs taux d'intérêts de prêts avec (2011, $111M). Dans la pratique, c'est une forme de système de micro-subprimes, 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 ! Dans la pratique, 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 !

Et enfin, l'Américain **H&R Block** 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.

Gestion financière d'entreprises

Ces 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** (2015, $3,9M) qui cible les TPE et PME.

Et puis de l'optimisation de planification financière d'entreprises avec **Anaplan** (2006, $239m), **Adaptive Planning** ($22,5m) et **Trufa** (2013, $15,9m).

Gestion des risques et 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, $53,6m) avec une offre de sécurisation généraliste destinée aux banques et commercants et aussi l'Israélien **Riskified** (2012, $63m), qui est focalisé sur les sites marchands.

La banque **Santander** utilise la reconnaissance vocale pour sécuriser les transactions, avec la solution de la startup espagnole **Fonetic** 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) qui aide à détecter les anomalies dans de nombreux marchés verticaux, dont la finance[162].

Deux startups américaines spécialisées dans les applications de gestion de la conformité des transactions **Lucid** (2004, $15M) et **Feedzai** (2009, $26M) utilisent toutes deux le machine learning pour détecter 80% des fraudes.

La gestion des risques porte aussi sur le credit rating[163] d'emprunteurs basé sur les informations disponibles sur les réseaux sociaux avec **TrustingSocial** (2015), qui n'est pas sans poser diverses questions sur le respect de la vie privée. De nombreux services de crowlending tels que **Kabbage** (2009, $488M), **Lending Club** (2007) et **ZestFinance** (2009, $67M) font aussi appel au machine learning pour le credit rating.

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, comme illustré *ci-dessous* [164]!

---

[162] Cf leur intéressant livre blanc Artificial Intelligence (AI) for Financial Services.

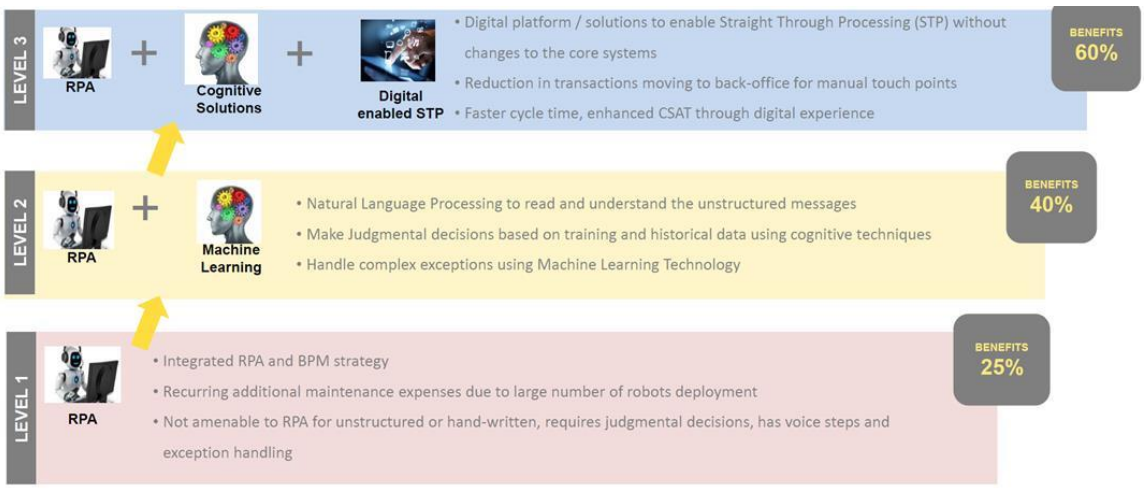
[163] Cf Application of Artificial Intelligence Techniques for Credit Risk Evaluation de Ahmad Ghodselahi et Ashkan Amirmadhi qui décrit une méthode de credit rating à base d'arbres de décision, de machine learning et deep learning exploitant une dizaine d'agents différents.

[164] Source : Robotics Process Automation 6 questions to master it, janvier 2017 de la société de conseil parisienne Ailancy.

Sector	Financial institution	Solution provider	Automation Type	Automation use
Insurance	Genworth	NA	Data migration & management	- Underwriting long term care and life insurance policies - A Logic rules engine backed by an algorithm encodes the underwriter guidelines - A natural language parser is also used to improve the coverage of the underwriting system
Wealth Management	UBS	NA	Support desk	- The robot replaces private bankers navigating through various applications and analysing client behaviours to provide adapted UBS products to wealthy clients and answering their requests. It frees up private bankers time for more added value activities - The robot models 85 millions singaporian individuals'behavioural patterns showing potential match-ups with various UBS products
Investment Banking	Goldman Sachs	KENSHO	Support desk	- Goldman established a strategic partnership with Kensho (US$15m investment) - Set up a platform for investment professionals that instantly and clearly answers millions of complex financial questions - This enabled speed and scale gains when automating a previously human intensive and time consuming task
	BBVA	fonetic	Data migration & management	- Trading record keeping to meet compliance requirements - The robot monitor internal communication and identify potential misconduct
	New York based Investment Bank	IPSOFT	Artificial intelligence	- The robot automates resolution of failed trades in fixed income securities - Due to complex and constant change in IT environment, 47mn were necessary to fix the frequent fails in trades - Within a month, 80% of failed trades are resolved without human intervention in 4mn and bank reached a 35% reduction in staff
Retail Banking	The co-operative bank	blueprism	Back Office Administration	- The robot improves speed and accuracy when dealing with customer queries on 10 process including direct debit cancelations and account closures - With the previous system, average audit took several hours and large one several days / Thanks to automation, it takes 1 to 30mn
	The co-operative bank	blueprism	Back Office Administration	- Returning or processing payments for clients with payments due and insufficient funds to meet obligations (based on account profiles of each customer / c.2,500 accounts daily) - 80% of the process is now automated with gains in consistency and speed
	New York based Investment Bank	IPSOFT	IT & Infrastructure support	- IT support staff are usually well trained and highly skilled and spend a lot of time in low added value tasks such as system checks and diagnostics, fault remediation, undertaking systems administration tasks and running identity checks when answering the clients - The robot acts as IT service desk agent providing more accurate answers and fastening the process
	New York based Investment Bank	IPSOFT	Support desk	- The robot answers mortgage brokers queries concerning banks'mortgage products - Usually, bankers use different systems and applications to answer brokers' questions. When they receive calls, they have to navigate through these applications when interacting with the broker on the call
	CAP GEMINI	AUTOMATION ANYWHERE	Data migration & management	- Automate tasks such as updating loan account data (entering values at the interface) , add reference data to optical files (pulling data from online reports), restore 'lost' teller receipts to backup files, etc. - Save the bank two work days each month and the automated process became much more accurate

25/01/2017

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[165].



Pas loin de nous, la startup roumaine **UiPath** (2012, $30M) est positionnée exactement sur ce créneau-là[166]. 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.

Cette forme de RPA reprend les anciens concepts des solutions de gestion de workflow.

---

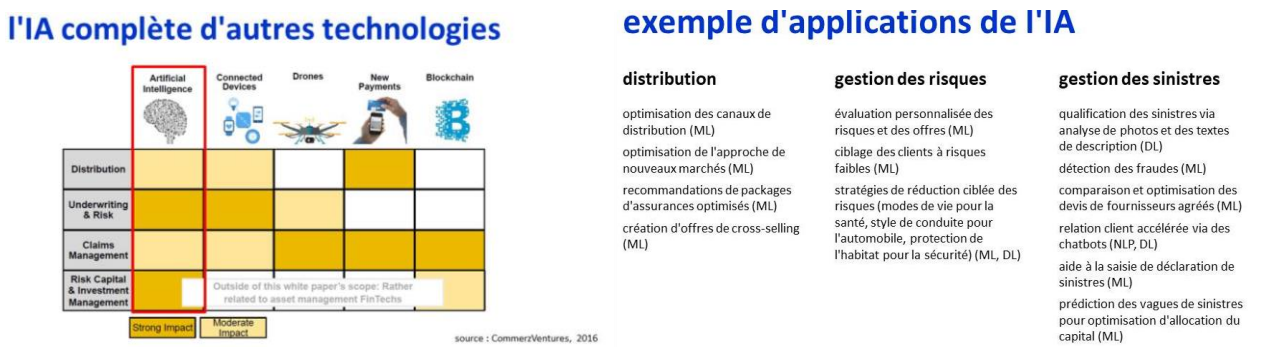
[165] Source : Introduction to Robotic Process Automation, a primer, de l'IRPA, Institute for Robotic Process Automation, 2015. Source du schéma : RPA and Beyond de TATA, juin 2017.

[166] Cf Les "robots logiciels" de cette startup roumaine prennent le travail peu qualifié d'employés de bureau, août 2017.

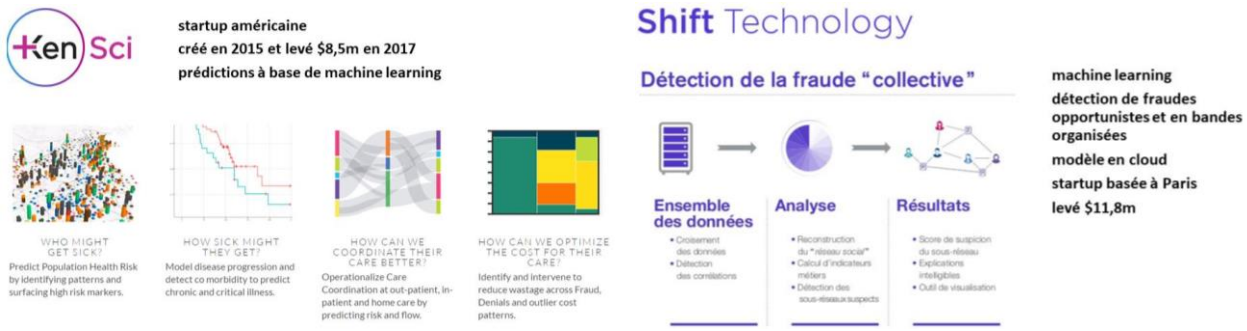
Assurance

Comme la finance, le métier d'assureur tourne essentiellement 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.

Dans l'assurance, l'IA intervient dans tout le cycle de vie : pour créer des produits, segmenter ses clients, cibler ceux qui présentent le moins de risques, proposer les bonnes offres aux clients, gérer le préventif, gérer les expertises et détecter les fraudes[167]. La relation client fait, comme dans les banques, aussi appel aux chat-bots[168].



De nombreuses startups se sont évidemment aussi lancées dans ce secteur et en exploitant des briques d'IA. Nous avons par exemple **KenSci** (2015, $8,5M) qui et spécialisé dans les prévisions à base de machine learning, et la française **Shift Technology** ($11,8M) qui est spécialisée dans la détection de fraude, avec une solution en cloud. **Riskgenius** est à l'origine d'une solution de traitement du langage qui gère le cycle de création et modification de contrats d'assurances. **Cyence** (2014, $40M) a créé une solution à base d'IA qui évalue le risques en cybersécurité d'entreprises clientes.



De son côté, l'assureur **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.

---

[167] Cf Emerging Technologies Transforming the $4tn Insurance Industry, de CommerzVentures, 2016.

[168] Cf https://chatbotsmagazine.com/why-chatbots-are-taking-over-the-insurance-industry-57b5151bb56a.

Les assurances font aussi appel à la reconnaissance d'image dans le cadre d'expertises, notamment automobiles ainsi que pour scanner les constats. L'anglais **Tractable** (2014, $10M) propose une solution d'inspection visuelle d'automobiles basée du deep learning pour de la classification automatique (vidéo).

Understory (2012, $9,7M) 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.

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.

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. Les assurances peuvent aussi encourager les conducteurs à faire auditer leur mode de conduite via des capteurs CANII dont l'offre est actuellement très abondante.

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, $52M) et à sa solution de gestion documentaire.

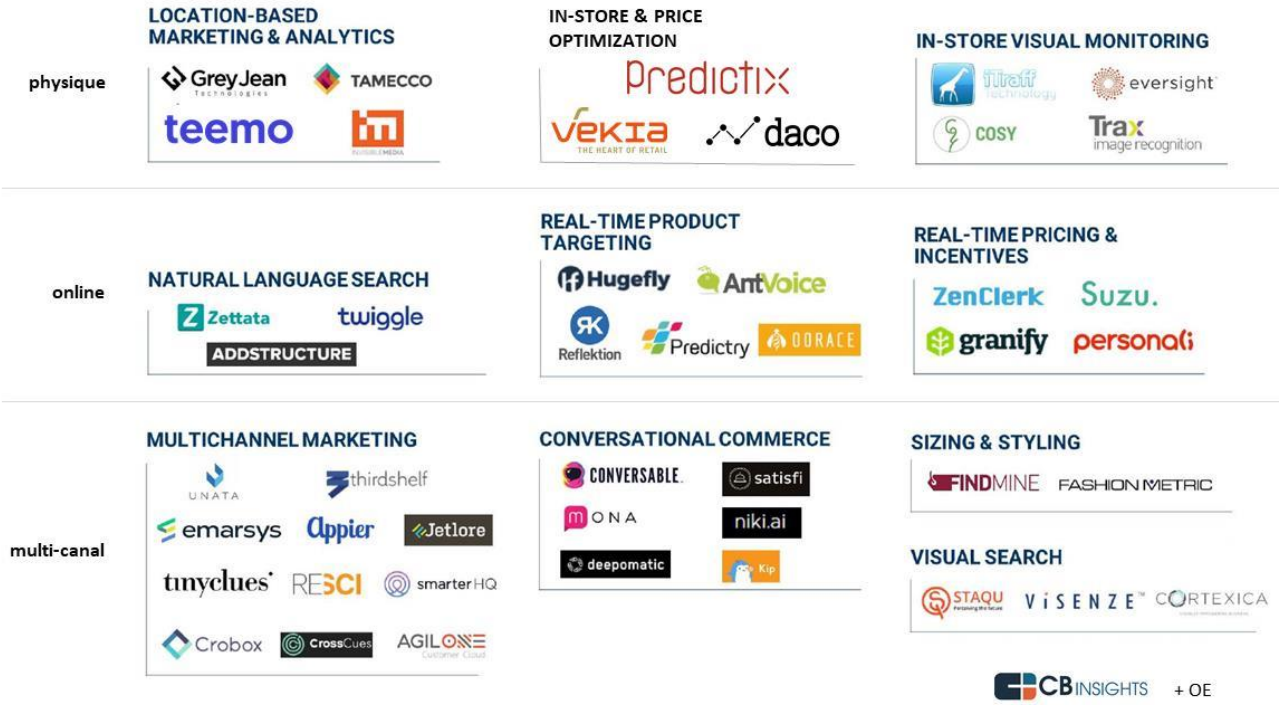
Enfin, une startup peut ambitionner de remplacer les compagnies d'assurance traditionnelles. C'est le cas de **Lemonade** (2015, $60M), est une société d'assurance en ligne pour propriétaires et locataires basée à New York qui s'appuie fortement sur l'IA dans tous ses processus et se passer d'intermédiaires (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, $25M).

Distribution

Le monde de la distribution présente des contrastes 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. 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.

Les besoins des commercants ? Ce sont des intermédiaires entre les marques et les consommateurs. Ils doivent analyse les tendances, comprendre les sentiments autour des marques, mettre les produits bien marketés dans les maisons des consommateurs, 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.

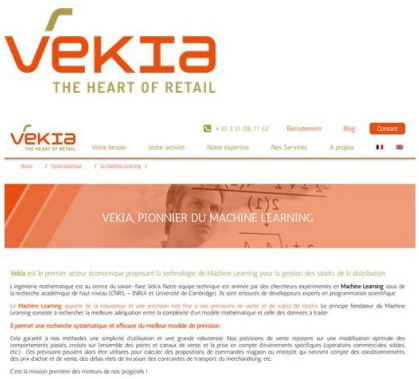


IA pour le retail physique

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.

Predictix (2005, $40M), qui fait partie d'Infor depuis 2016, est spécialisé dans l'optimisation de linéaires. Comme son nom l'indique, il doit utiliser des techniques de machine learning pour faire du prédictif. **Celect** (2013, $15,3M) permet aussi d'optimiser l'assortiment des rayons en fonction d'analyses prédictives comportementales des clients.

Le français **Vekia** (2008, $2,7M) utilise le machine learning pour faciliter l'optimisation d'approvisionnement. Sa solution est déployée chez Leroy Merlin, But, MrBricolage et Jacadi.



startup de Lille

optimisation d'approvisionnement

chez Leroy Merlin, But, Mr Bricolage, Jacadi

augmente de 4% le CA du linéaire

L'Américain **Reflektion** (2012, $27,8M) 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.

Le français **Teemo** (2014, $17,6M), 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, Carrefour et Casino.

La startup française **Occi** (2015) 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.

"Real Life Targeting"

cible les mobinautes en fonction des lieux qu'ils ont visités et mesure l'impact sur le trafic généré en magasin.

>100 clients dans l'automobile, l'alimentaire, le bricolage et l'ameublement (Volkswagen, Carrefour, Brico Dépôt, Gautier)

startup française, ex Databerries

50 personnes à Paris, Londres et New York.

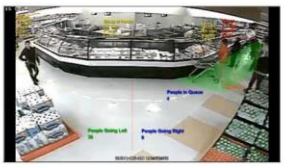
levé $17,6m

Neosensys

optimisation de linéaire

algorithmes originaires de l'Inria

exploite vidéos des caméras de surveillance



Opius

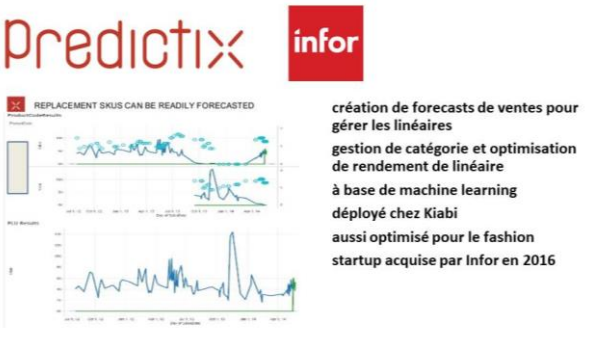
"deep learning" pour reconnaître les produits en magasin à partir de photos

chez 50Partners

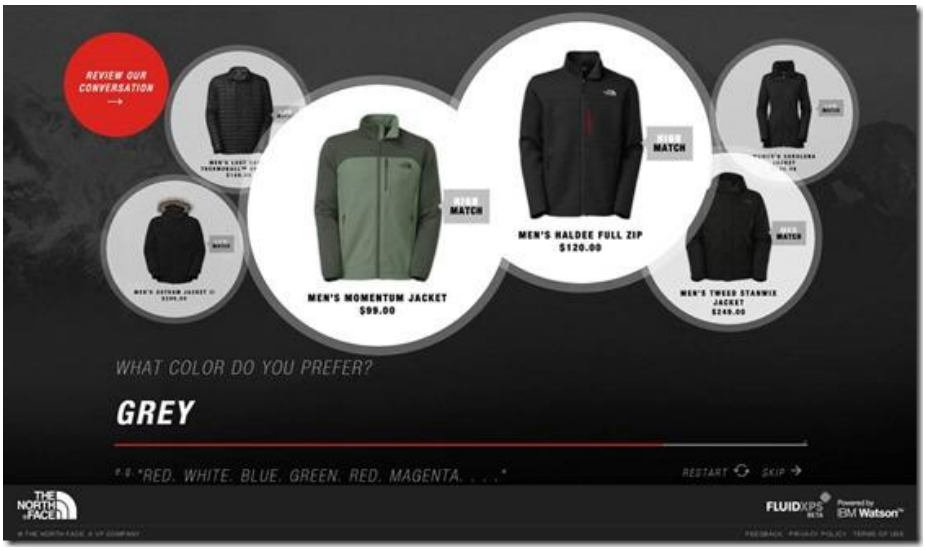
Le français **Neosensys** est un spécialiste de la vidéo-surveillance avec deux cas d'usage : la détection et la prévention de la démarque inconnue grâce à un suivi multi-caméras de personnes suspectes et l'optimisation de linéaire en fonction de l'analyse des parcours des clients dans les points de vente. La société collabore avec l'INRIA pour le développement de ses solutions.

Percolata (2012, $5M) 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.

Quidivi (2006, $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.



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, $24M). 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[169].

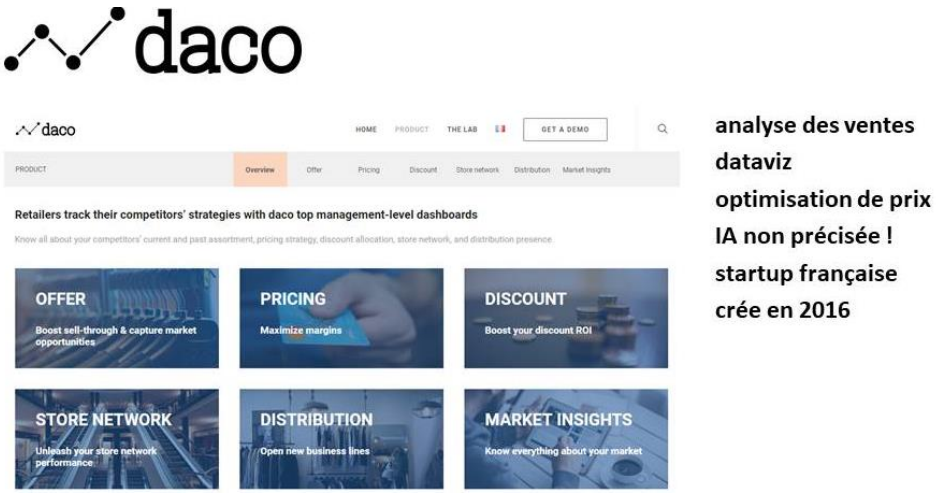
IA pour le commerce en ligne

Le commerce en ligne intègre d'abord presque tout ce que l'on trouve de nouveau dans le marketing, pour le ciblage publicitaire et commercial et sur les techniques d'upselling et cross-selling basées sur la recommandation qui s'appuient sur le machine learning.

Le commerce en ligne peut exploiter d'autres nouveautés à base d'IA :

[169] Pour en savoir plus voir ce compte-rendu détaillé sur le JDN : **Comment The North Face a appliqué Watson à l'expérience d'achat**.

- L'optimisation de l'activité de commerciaux sédentaires et prévision du comportement des clients avec des startups particulièrement bien financées : **InsideSales** (2004, $251M), **Lattice** (2015, $9,3M), **Clari** (2012, $26M), **Wise.io** (2012, $3,6M) et **Spiro** (2014, $1,5M). Il est assez difficile de départager toutes ces startups !
- L'optimisation du parcours client en ligne avec **Gainsight** (2011, $154M), **Jetlore** (2011, $7M) et **OnCorps** (2011, $2,3M). Le canadien **Granify** (2011, $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. Le français **Influans** (2016, $6M) optimise aussi le parcours client pour leur proposer le bon produit avec les bonnes incitations et au bon moment. La startup a été créée par deux anciens fondateurs de Talend, Bertrand Diard et Fabrice Bonan.
- **TargetToSell** (2012, 5M€) est une autre startup française qui optimise le parcours client au niveau du site et pour chaque visiteur. La startup 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.
- De nombreuses solutions permettant de trouver la bonne taille et le bon style pour s'habiller avec l'anglais **Thread** (2012, $16,32M), le styliste en ligne **Stitch Fix** (2011, $42M) qui exploite le dialogue en langage naturel, **Volumental** (2012, $3M) pour le choix de ses chaussures et **Thirdlove** (2013, $13,6M) pour la taille de son soutien-gorge.
- L'optimisation de prix en fonction de l'activité des concurrents, avec le français **Daco** (2016) qui s'appuie sur du deep learning.



- Différents moteurs de recherche avec l'israélien **Twiggle** (2014, $35M) qui permet 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, $14M), **Cortexica** ($6,6m de levés) et son logiciel findSimilar en cloud, **Grokstyle** (2016, $2M) et **Slyce** ($37m de levés et

IPO en avril 2015). Et puis le moteur de recherche FashionBot de **GoFind** (2016) 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

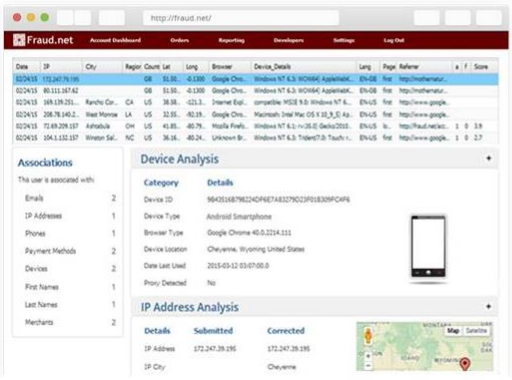
API en cloud

startup isrélienne créée en 2014

$35m levés

- Un système de recommandation basé sur du machine learning avec **Dato** (2013, $23,5m), 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, $27M), adopté par Disney et Converse, qui propose du ciblage produit temps réel.
- La détection de fraude pour le commerce en ligne avec **Fraud.net** qui s'appuie probablement sur du machine learning.

Fraud.net



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

startup de New York créée en 2013

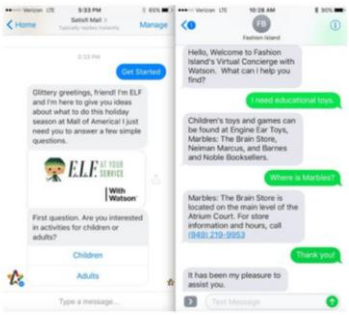
- Des chatbots en tout genre avec l'américain **Satisfi Labs** (2016) et l'agence française **TheChatbotFactory** qui créé des chatbots sur mesure. Elle a créé un chatbot sommelier pour Auchan sous Facebook Messenger.

THECHATBOTFACTORY

chatbot sommelier chez Auchan en test sous Facebook Messenger

satisfi

chatbot réalisé avec IBM Watson utilisé notamment chez Macy's



Les ecommancants 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€[170]. Et Amazon est le premier client en vue pour ce genre de solution.

Médias

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 entraine 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[171] et si cela prête à conséquence. L'IA est comme Internet et Google

[170] Nombreux details ici : http://www.businessinsider.fr/us/mercedes-electric-vision-van-drone-delivery-service-photos-2017-3/.

[171] Cf http://www.meta-media.fr/2017/05/07/et-si-les-medias-redevenaient-intelligents.html.

Search. Elle peut entrainer 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 reconnaitre des personnes dans des photos, ce qui est très utile pour comprendre des situations captées par des photo-journalistes.

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), un outil d'analyse de contenus qui permet de raliser des recherches, des liens entre documents, de les classifier et de les exploiter, le tout sur plusieurs sources et plusieurs langues.

Production

Nombre de startups sont apparues ces dernières années qui automatisent la création 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.

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 faible partie d'entre elles ciblent les médias, pour ce que l'on appelle le « robot journalism » comme le français **Syllabs** et sa solution Syllabs Media ou encore l'Américain **Textomatic** (2010, $40M). La raison est simple : ce marché est moins solvable que la finance ou le 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)[172]. L'article n'a dans la pratique 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.

[172] 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.

SECTIONS SEARCH Los Angeles Times SUBSCRIBE 4 weeks for 99¢ LOG IN

EDITION: CALIFORNIA | U.S. & WORLD LOCAL SPORTS POLITICS ENTERTAINMENT OPINION MOST POPULAR PLACE AN AD THURSDAY AUG. 24, 2017 68°

3.8 earthquake shakes Los Angeles area

SHARE THIS A shallow magnitude 3.8 earthquake was reported one mile from View Park-Windsor Hills

MAY 3, 2015, 10:25 AM

A shallow magnitude 3.8 earthquake was reported Sunday morning one mile from View Park-Windsor Hills, according to the U.S. Geological Survey. The temblor occurred at 4:07 a.m. PDT at a depth of 5.6 miles.

The quake was classified by the USGS as "light" but was felt over a wide area of the L.A. basin. The Los Angeles Fire Department said it had received no reports of damage.

RELATED: Baldwin Hills-area quakes not linked to oil operations, experts say

A 3.5 quake rattled the same general area on April 12. Both quakes were centered on the Baldwin Hills/Inglewood border. The Newport-Inglewood fault runs along that area.

In the aftermath of that quake, some residents asked whether oil production in the area might have been a factor. But USGS seismologist Lucy Jones wrote on Twitter that it was unlikely because the depth of the quake was so far below oil production facilities.

Staff writer Shelby Grad contributed to this report.

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, $650K) 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-dessus à droite*). Elle ajoute des métadonnées aux scènes analy-

sées exploitables dans les outils d'analytics voire pour les générateurs de guides de programmes.

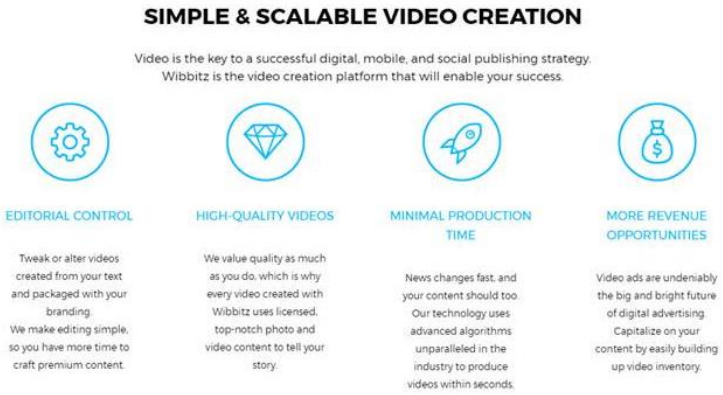
Le breton **Mediawen** (2014) gère la traduction de contenus vidéo en temps réel en s'appuyant sur IBM Watson puis text to speech, en voix de synthèse ou sous titrage.

WIBBITZ

génération automatique de vidéos d'actualités à partir de contenus textuels

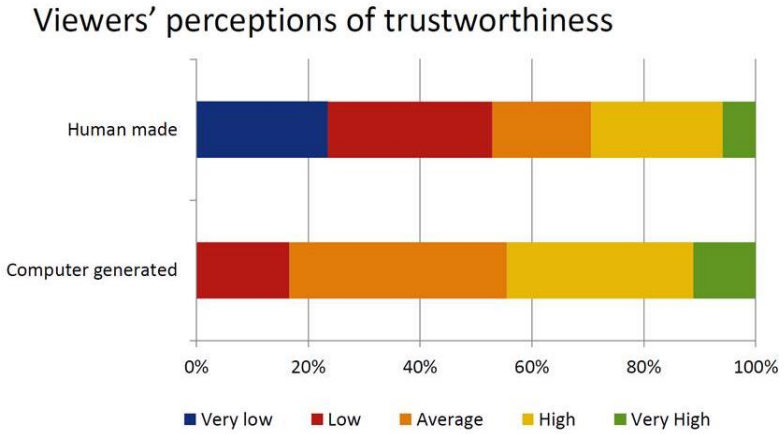
startup israélienne créée en 2011

levé $11,8m



L'israélien **Wibbitz** (2011, $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.

Une étude américaine montre que les lecteurs font plus confiance à des articles produits par des robots que par des journalistes[173]. 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[174] !



[173] Source du schéma à droite : The Artefacts of Automated Journalism: Producers' Perspectives and Audience Assessments de Neil Thurman.

[174] Cf http://www.newstatesman.com/science-tech/technology/2017/03/human-journalists-hate-robot-journalists-says-new-report.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 !

Mais 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 vidéos ! Bref, le pire ! 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 !

L'industrie musicale tire aussi parti de l'IA pour la production et la diffusion de contenus.

- Le canadien **Landr** (2012, $9m) propose un service en cloud d'automatisation du mixage audio, qui va créer des morceaux de musique agréables à l'écoute (vidéo).
- L'Américain **Popgun.ai** (2017) 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, $4M) 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.
- **Pacemaker** (2011, $4M) est un DJ à base d'AI qui exploite les contenus de services de musique en ligne comme Spotify.

Diffusion

Newswhip (2011, $9M) 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. C'est aussi ce que fait **Banjo** (2010, $121M) et qui ne cible pas que le marché des médias, sinon, il n'aurait pas levé ce montant.

NEWSWHIP

analytics de l'écho des médias et sujets dans les réseaux sociaux

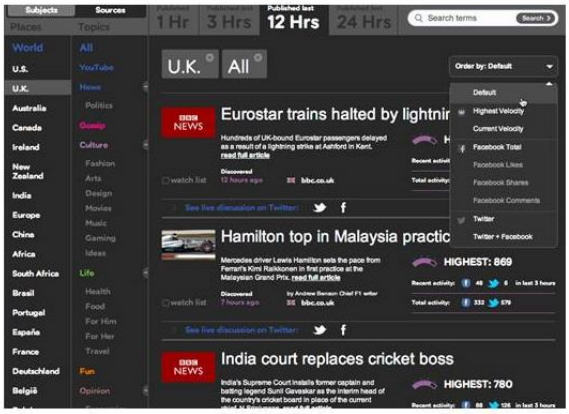
réalise des prédictions de sujets à fort écho

à base de machine learning

références : Huffington Post, BuzzFeed, BBC, The Guardian

startup de New York créée en 2011

levé $9m



La startup **Echobox** (2013, $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 !

Enfin, la start-up de découverte de musique **Decibel Music Systems** (2010) utilise IBM Watson dans son application MusicGeek pour faire de la recommandation.

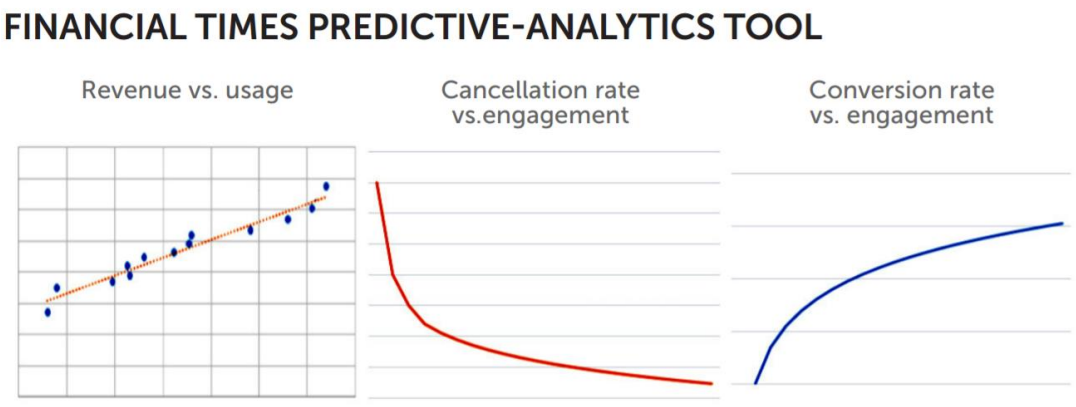
Monétisation

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, $4,2M) 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. 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, $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[175].



The Financial Times predictive analytics tools correlate data about revenue, content usage, cancellation rates, engagement, and conversion.

[175] Source : https://digitalcontentnext.org/blog/2017/06/13/artificial-intelligence-gains-momentum-news-media/.

Tourisme

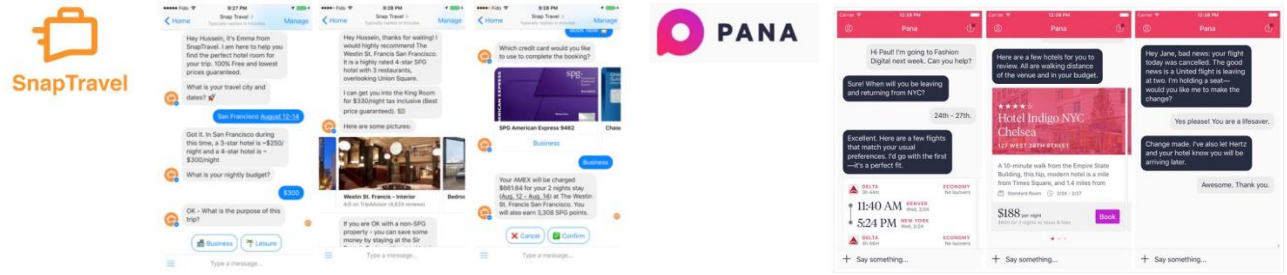
Le tourisme est un autre terrain de jeu propice aux innovations à base d'IA : les données sont abondantes, notamment via les intermédiaires de la réservation en ligne et chez les compagnies aériennes et leurs services de réservation mutualisés tels que Sabre et Amadeus, c'est un marché grand public, il peut exploiter les outils de la mobilité et ceux des objets connectés.

Les systèmes de réservations de vols et hôtels exploitent toutes les techniques imaginables de « yield management » pour les remplir aux prix les plus élevés. Certains systèmes exploitent de la logique floue.

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**[176], **Air France KLM** qui a un chatbot pour Facebook Messenger, avec le chatbot de **Voyages SNCF** sur Facebook Messenger (vidéo), **Mercure Bot,** toujours sur Facebook Messenger[177] ou **Ask Mona**, un chatbot de sélection de visites culturelles en France.

SnapTravel (2016, $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. Bref, c'est un moteur de recherche à commande textuelle. **Pana** (2015, $1,45M) est un équivalent destiné aux voyages professionnels.



Les applications mobiles généralistes chatbot de préparation de voyage ne manquent pas aux USA. Nous avons par exemple **Lola** (2015, $44,6M), créé par Paul English, le fondateur du moteur de recherche de voyages **Kayak** revendu à Priceline pour $2B**, Mezi** (2015, $11,8M) et **Skyscanner** (2003, $197M, acquis par le chinois Ctrip en novembre 2016). Toutes ces applications se ressemblent et accèdent généralement aux mêmes sources de données[178].

---

[176] Voir cette liste de quelques dizaines de chatbots : https://www.30secondstofly.com/ai-software/ultimate-travel-bot-list.

[177] Que j'ai testé et qui n'apporte pas grand-chose, et en plus est très lent.

[178] Cf http://www.nanalyze.com/2017/04/artificial-intelligence-ai-travel/.

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.

L'Américain **Wayblazer** (2014, $5M) 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 l'Américain **GoMoment** (2010).

Gogobot (2010, $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**.

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

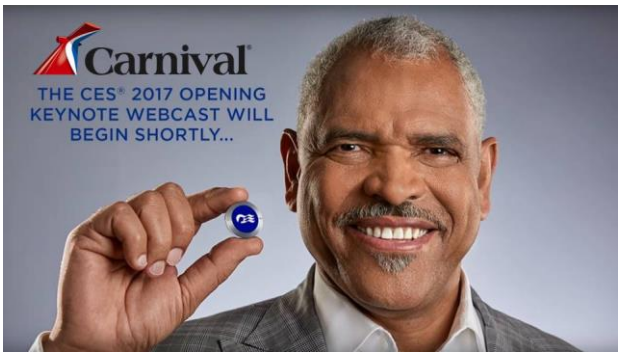
L'expérience touristique peut s'améliorer en tirant parti de l'IA à différents étages.

J'ai pour l'instant découvert cet outil de prévision proposé aux hôteliers par la startup française **Victor&Charles** qui 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.

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.

Et puis surtout, cette expérience client en environnement fermé présentée par **Carnival** lors d'un keynote au CES 2017 (vidéo). Elle consiste à proposer un badget 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.



Robots

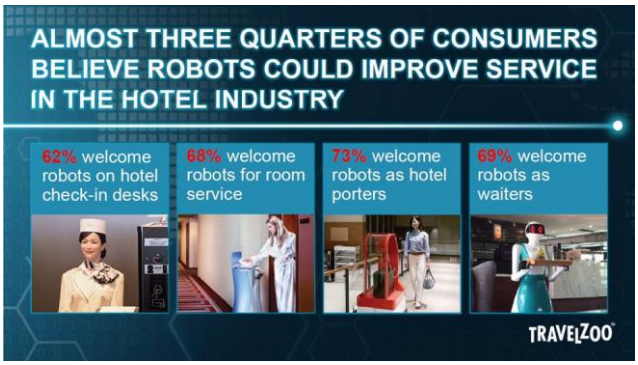
Si vous aimez les robots, le tourisme pourra vous donner l'occasion d'en croiser.

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.

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.



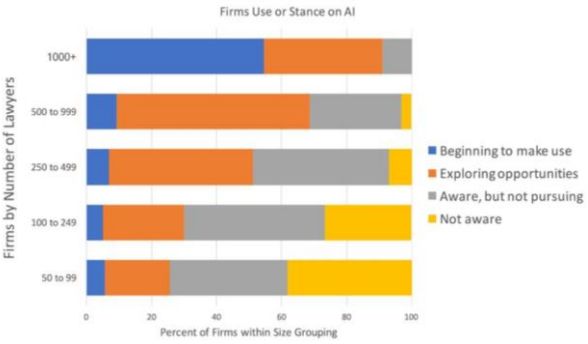
Ce n'est qu'un début ! Après, si vous avez besoin de rencontrer de vraies personnes pour alimenter votre organisme en sérotonine et en dopamine, c'est une autre affaire !

Juridique

Le lancement de la startup canadienne **Ross Intelligence** (2014, $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[179].

Qu'en est-il dans la pratique ? Quand on y regarde de près, on constate qu'une bonne partie des outils de l'IA dans les métiers juridiques sont des moteurs de recherche améliorés qui permettent de consulter les lois et la jurisprudence.

En complément, des applications plus élaborées voient le jour, pour comparer des contrats, pour les optimiser et pour faire aussi des prévisions sur l'issue d'un procès. Dans l'ensemble, les techniques d'IA exploitées dans le domaine juridique tournent autour du NLP (Natural Language Processing).



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 utilisent 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.

[179] Cf 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.

En même temps, les solutions d'IA juridiques ne vont pas remplacer les avocats et juristes qui devront encore longtemps gérer leurs clients et plaider. Elles affecteront surtout les métiers du paralegal dans les cabinets. Elles amélioreront 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 les prix.

Dans **Legal Aspects of Artificial Intelligence**[180] 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. A chaque fois, il s'agit d'effets d'annonce, sans qu'il y ait encore de véritables retours d'expérience.

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
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]

Moteurs de recherche juridiques

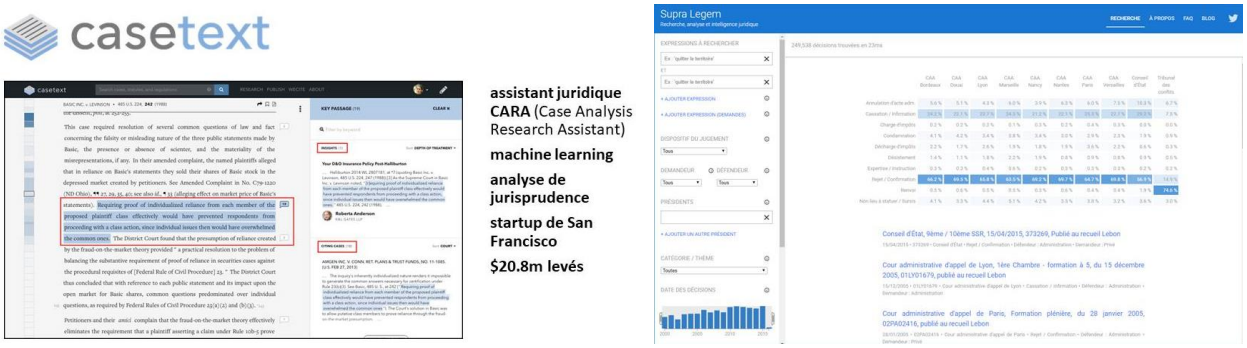
Une grande majorité des startups juridiques de l'IA proposent donc des moteurs de recherche dans les textes de lois et de jurisprudence. Elles florissent particulièrement bien dans les pays anglo-saxons dont le droit est influencé par la jurisprudence (dit de « case law »), tandis que le droit européen et surtout français, est plus fortement influencé par les lois et règlements (« common law », ou droit romain).

Ross Intelligence (2014, $120K), 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 par quelques grands cabinets d'avocats aux USA comme **Baker Hostetler** en mai 2016. 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

[180] Cf https://www.lexology.com/library/document.ashx?g=53ef5f3a-b41a-4a24-ba7c-1ec6d53187f7.

qui n'est pas évident à comprendre car la réponse à des questions ne constitue pas une forme d'apprentissage supervisé ou par renforcement.

Nous avons divers autres systèmes d'interrogation de bases de connaissances comme **Ravel Law** (2012, $15M), issu de l'école de droit et celle d'informatique de Stanford et acquise par LegalNexis en 2017, **Casetext** (2013, $20,8m) qui est focalisé sur l'analyse de jurisprudences et **Judicata** (2012, $7,8m) qui lui aussi s'intéresse à l'analyse de la jurisprudence.

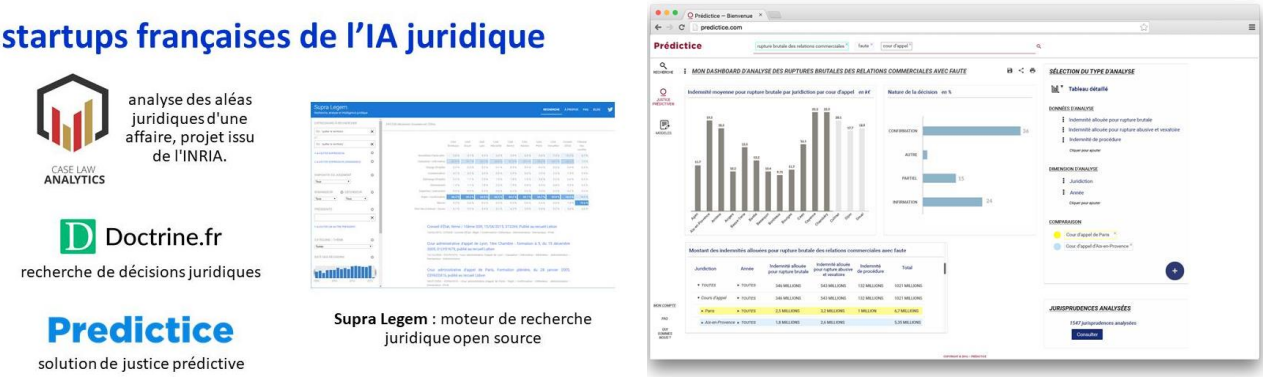


En France, **Supra Legem** est une ressource en open source et open data pour accéder à la jurisprudence (*ci-dessus à droite*).

Simulations et prévisions

Le français **Case Law Analytics** est une startup proposant une solution d'analyse des aléas juridiques d'une affaire. Là encore, il doit y avoir beaucoup de NLP 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 analysées. Si votre cas est inédit, l'étude de la jurisprudence n'apporte pas grand-chose. La startup qui est portée par l'INRIA est en phase d'amorçage.

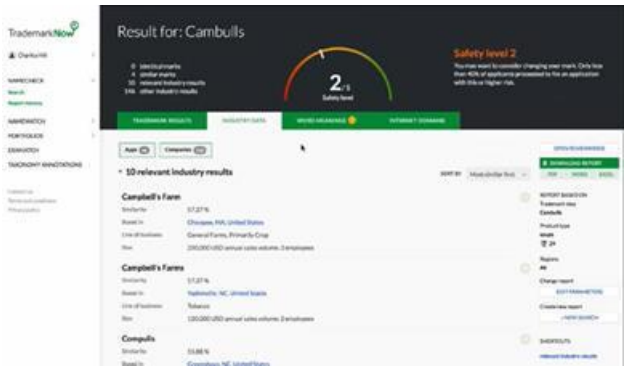
Un autre français, **Predictice** (2016), est aussi positionné sur la justice prédictive. C'est cependant une solution généraliste avec un moteur de recherche de documents juridiques. La startup exploite les données ouvertes de Légifrance (textes de droit) et Jurica (jurisprudence). Au passage, elle est hébergée chez OVH.



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 étant spécifiques, celles des marques et des brevets.

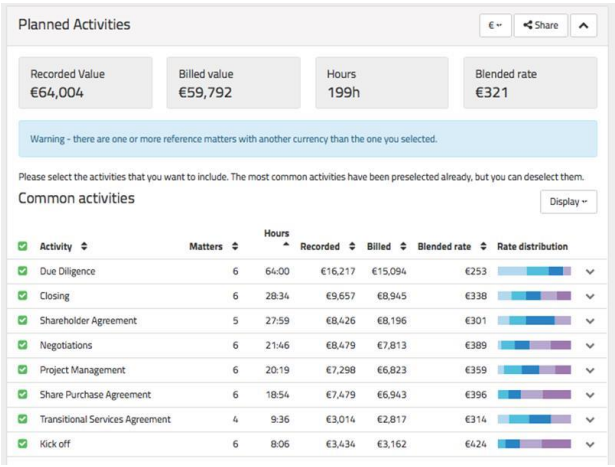
- **Juristat** (2012, $1,62M) réalise des analytics sur les données publiques sur les brevets et les avis des agents de l'USPTO pour optimiser les stratégies de protection de brevets. C'est une solution dédiée au marché US. Le financement de la startup l'explique en partie. Il est encore limité pour permettre un développement international rapide.
- **Turbo Patent** (2008, $3,45M) produit des reportings sur la qualité et la valeur d'un portefeuille de brevets.
- **Lex Machina** (2009, $10M) 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.
- **Onomatics** (2012, $3,5M) propose TrademarkNow qui est focalisé comme son nom l'indique sur les recherches portant sur la protection des marques (*ci-dessous*). Il est cependant difficile d'y identifier des morceaux d'IA.



Autres usages

Business analytics, analyse de contrats, gestion de contrats de confidentialité et chatbots grand public au programme !

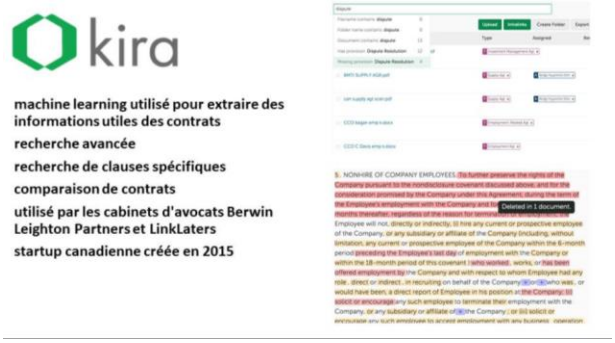
ClockTimizer est un outil de business analytics pour cabinets d'avocats qui permet d'évaluer le temps passé sur des contrats clients et d'affiner ensuite les devis pour d'autres clients basés sur l'expérience. Les outils exploitent à la fois des données textuelles (mots clés des contrats, etc) et quantitatives (temps passé, etc).



Kira Systems (2015) propose une solution de « due diligence » et d'analyse de contrats. C'est aussi l'activité d'un autre Canadien, **eBrevia** (2012, $2M) qui a codéveloppé sa solution avec l'Université de Columbia qui couvre notamment le droit de l'immobilier.

Neota Logic est une startup qui est spécialisée 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.

DoNotPay est un chatbot anglais créé par un jeune de 19 ans qui permet de faire sauter les contraventions aux USA et au Royaume Uni. 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.



On compte notamment **LegalZoom,** un service d'avocat en ligne couvrant à la fois le droit des affaires et le droit civil[181] qui s'appuie sur IBM Watson.

[181] Voir ce numéro de The American Lawyer qui en parle bien.

Services publics

L'IA peut servir de manière générique dans les services publics, notamment pour la création de chatbots, pour l'analyse de données avec du machine learning, pour mener des analyses macro-économiques et étudier les données de recensement.

Du côté de l'éducation, les applications de l'IA ne sont pas très nombreuses en l'état. Elle intervient dans les **universités** pour les aider à recruter les meilleurs étudiants **Plexuss** (2014) ou, au contraire, pour aider ces derniers à trouver la meilleure université avec **Admitster** (2014). Mais on parle des USA, pas de la France !

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**[182].

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[183]. 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[184].

---

[182] 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 ! Cf 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.

[183] Cf China seeks glimpse of citizens' future with crime-predicting AI en juillet 2017.

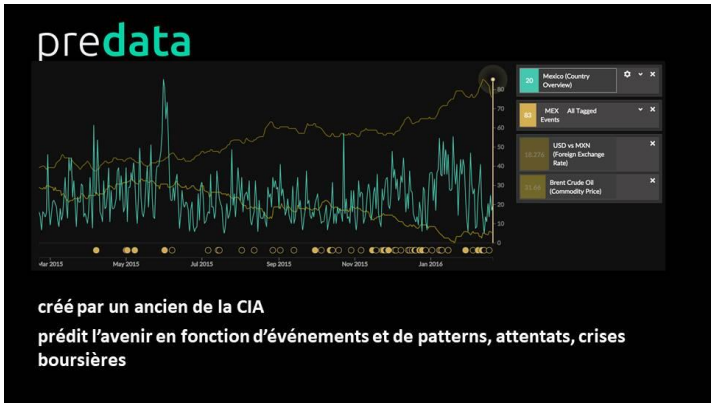
[184] Encore faut-il alors intervenir ! On n'a pas besoin d'IA pour identifier les pick-pockets dans les lieux touristiques à Paris et ils ne sont pas arrêtés pour autant.

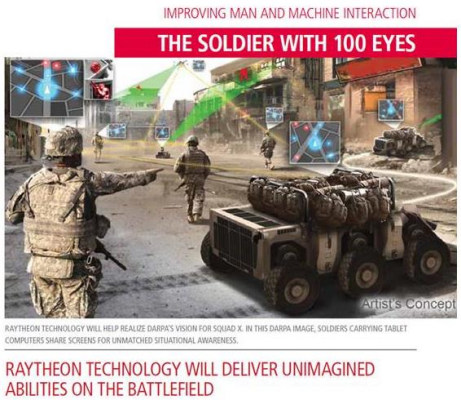
Renseignement et défense

L'IA a évidemment de nombreuses applications dans le vaste secteur de la défense et du renseignement.

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 télécommandé) et surtout avec les drones aériens, eux-aussi télécommandés mais dotés d'outils de reconnaissance de leur environnement et de pilotage automatique.

L'IA est aussi utilisée dans la prévision d'événements comme chez **Predata** (2015, $3,25M).





Elle peut 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.

Acteurs de l'intelligence artificielle

Nous allons dans cette partie décrire une bonne part de l'écosystème de l'IA : avec les grandes entreprises (américaines) du secteur, les startups et leurs tendances, et ce qui se passe en France aussi bien côté recherche que startups.

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. 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 Nathanel Rochester, un chercheur d'IBM. Et Microsoft a créé son laboratoire de recherche en 1991, principalement dédié aux avancées de l'IA.

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. La réussite dans l'IA comprend quelques ingrédients de base : une plateforme logicielle intégrée et ouverte, des ressources en cloud éventuellement enrichies par des architectures matériels propriétaires différenciées, l'accès à des données sachant que les plus intéressantes proviennent d'activités grand public (ci-dessous, la liste des aplications de leaders, y compris les chinois Alibaba et Baidu), des partenaires adoptant la plateforme, ou la capacité interne à gérer des projets clients et enfin, surtout, des talents.

Sur les grands marchés horizontaux, ces GAFAMI et leurs équivalents chinois ont de fortes chances de devenir les leaders de l'IA[185]. Il reste probablement de la place pour des acteurs positionnés sur des marchés verticaux que ces grands acteurs ciblent mal, même IBM.

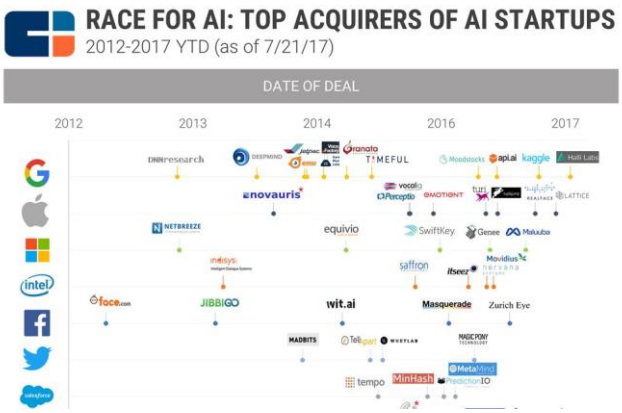
[185] Cf Why AI consolidation will create the worst monopoly in US history, août 2016.

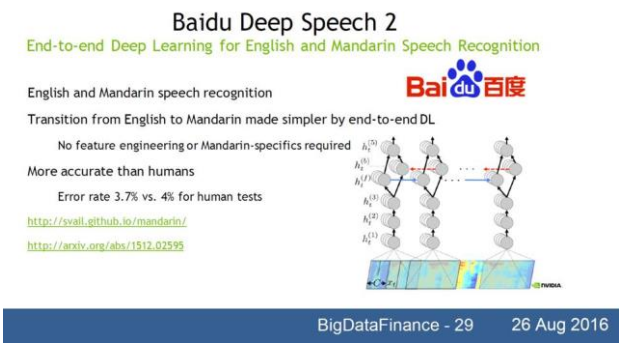
Est-ce que les leaders chinois vont dépasser les GAFAMI ? C'est une hypothèse à la mode[186]. Les grandes entreprises chinoises du numérique[187] et les startups chinoises bénéficient d'un terrain favorable : une masse de chercheurs et développeurs formés en Chine ou dans le reste du monde, un marché intérieur de plus de 500 millions d'Internautes, un marché mobile ultra-développé et une règlementation qui se pose beaucoup moins de questions sur la protection de la vie privée.

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. Si Baidu et d'autres font quelques progrès significatifs dans le deep learning, 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[188].

Ces grands leaders chinois ont recruté des talents chez des concurrents américains. 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. La recherche chinoise fait des progrès mais est encore en « piggy-back » d'une bonne part de la recherche nord-américaine (USA + Canada).

Bref, comparer la Chine aux USA en matière d'IA n'est pas évident, surtout si les indicateurs sont anecdotiques et pas économiques.





[186] Cf China May Soon Surpass America on the Artificial Intelligence Battlefield, février 2017 et China's AI awakening, paru en octobre 2017.

[187] 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.

[188] 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 tels système est liée à la diversité des visages qui servent à son entrainement. 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 entrainer 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.

L'autre benchmark habituel des grands groupes 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.

Les GAFAMI adoptent 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 (comme dans le hardware) ou sa dimension d'usages (pour le grand public).

Cette intégration verticale couvre de nombreux domaines comme les chipsets neuromorphiques (IBM, Google, Microsoft), les serveurs (Facebook, IBM), 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).

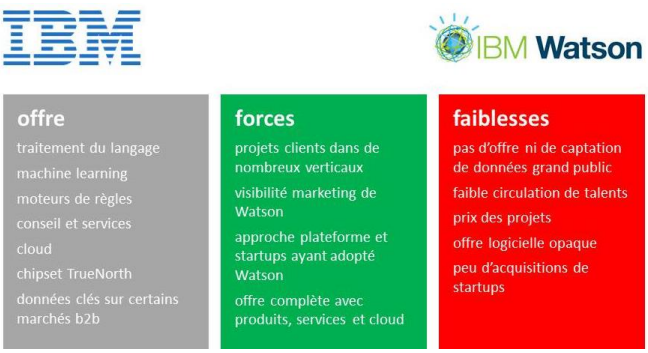
Pour résumer la situation des GAFAMI :

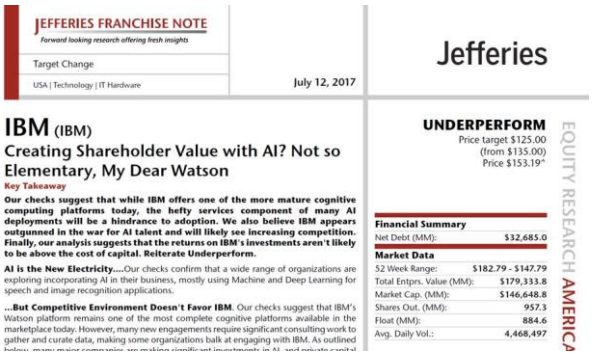
- **IBM** est un acteur très présent dans les grandes entreprises avec sa plateforme logicielle « couteau suisse » IBM Watson et son activité de services. C'est le grand du secteur qui est investi dans l'IA depuis ses débuts, en 1955 !
- **Google** est un acteur dominant du numérique grand public couvrant de larges pans de la vie des utilisateurs (recherche, email, mobile, TV, maison connectée). Cela lui permet de capter de gros volumes de données pour alimenter « ses IA ». Il domine aussi le développement de solutions d'IA avec son framework open source TensorFlow.
- **Amazon** est le leader du retail en ligne et du cloud d'entreprise. Il s'est taillé une bonne place dans les chatbots vocaux avec Alexa.
- **Microsoft** propose une belle plateforme logicielle d'IA couvrant le machine learning et le deep learning, ainsi qu'une excellente activité de recherche fondamentale, mais l'éditeur souffre d'un déficit marketing dans le domaine, notamment auprès des startups.
- **Facebook** domine les réseaux sociaux et la communication mobile (WhatsApp, Instagram), des logiciels qui exploitent de plus en plus d'IA. Il domine le marché des chatbots avec Messenger.
- **Apple** est une société très intégrée verticalement dont l'approche plateforme, surtout en cloud, est moins prégnante dans l'industrie. Mais iOS étant un passage obligé, il est mécaniquement présent chez de nombreux développeurs.
- **Oracle**, **SAP** et **SalesForces** ont mis l'IA à leur menu mais font moins parler d'eux du fait de leur positionnement entreprise.

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[189] décrit bien les enjeux d'IBM vus par la lorgnette des actionnaires : Watson est un bel 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. Elle fait face à des concurrences multiformes : les grandes sociétés de services et intégrateurs (CapGemini, ATOS, Orange Business Services pour la France), les grands éditeurs de plateformes (Microsoft, Oracle, …) et les fournisseurs de ressources en cloud (Amazon en tête). Qui plus est, 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.

Google et Facebook sont des concurrents plus indirects, ceux-ci n'ayant pas de véritable approche des grandes entreprises pour leurs projets stratégiques.

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 qui avaient aussi peur de perdre leur poste.

Fast forward. IBM a du faire 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.

[189] Cf IBM Creating Shareholder Value with AI? Not so Elementary, My Dear Watson, Jefferies Franchise Note, juillet 2017.

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 "one-stop-shop" provider d'IBM 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.

La première incartade d'IBM dans l'IA s'est manifestée au grand jour avec la victoire de l'ordinateur IBM 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**.

Dans partie intéressante et moins médiatisée[190] 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 disponibles sur Wikipedia.

Il ne sait pas non plus identifier des recettes de cuisine en fonction de leurs composantes. Watson a 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.

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, comme passer le 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 totalisant 16 To de RAM.

Watson est devenu une plate-forme logicielle, respectant en cela les canons de la réussite dans le numérique. Elle est proposée aux développeurs sous forme d'APIs en cloud. L'histoire est bien **racontée ici**.

[190] Cf https://www.youtube.com/watch?v=YgYSv2KSyWg.

Dans la pratique, 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.

IBM avait annoncé investir plus de $1B sur le Cognitive Computing, un peu comme il avait annoncé au début des années 2000 investir la même somme sur le développement de Linux. C'est donc un beau pari marketing et business qu'IBM fait ici. Et c'est plutôt bien vu car une bonne part du futur des solutions numériques va utiliser les techniques de l'IA. Il faut toujours se positionner sur un futur pas trop lointain pour éviter de rater les trains de la technologie qui passent !

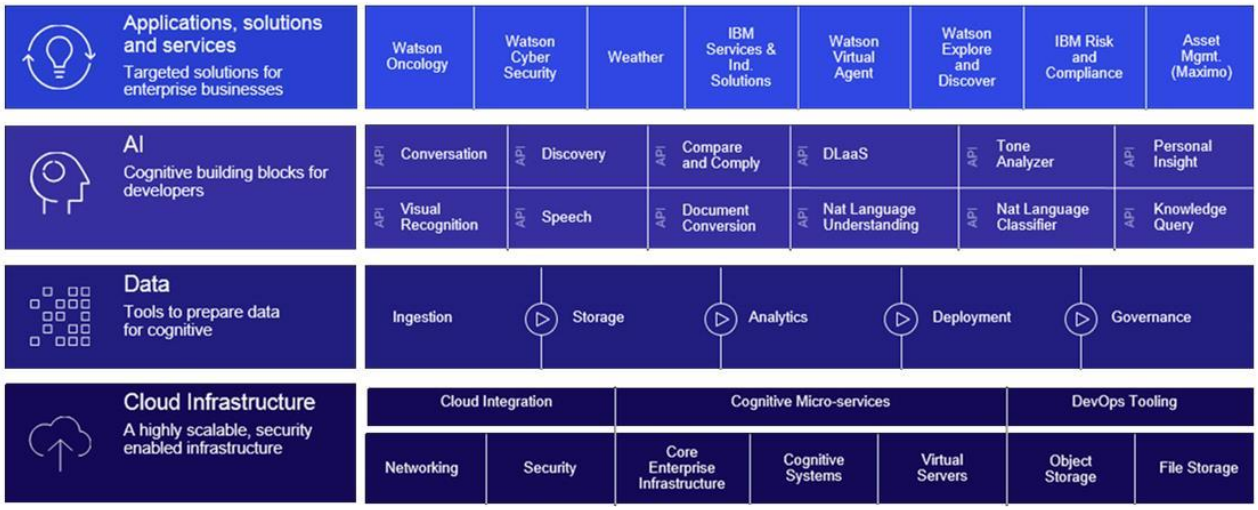
Depuis 2011, Watson est devenu le sujet phare de la communication d'IBM. C'était d'ailleurs le seul thème de l'intervention en **keynote** de Ginni Rometty au CES de Las Vegas de janvier 2016. IBM organise aussi chaque année une grande conférence "World of Watson" dont la dernière édition avait lieu à New York du 23 au 24 mai 2016[191] et l'édition parisienne avait lieu le 10 octobre 2017.

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 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[192] : dès que l'IA peut jouer un rôle dans un projet, « il peut le faire ».

IBM Watson n'est pas un produit. C'est une architecture et une plateforme faite de nombreuses briques logicielles.



[191] Les vidéos des keynotes de l'édition de mai 2015 sont disponibles pour la **première** et la seconde journée.

[192] Cf 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.

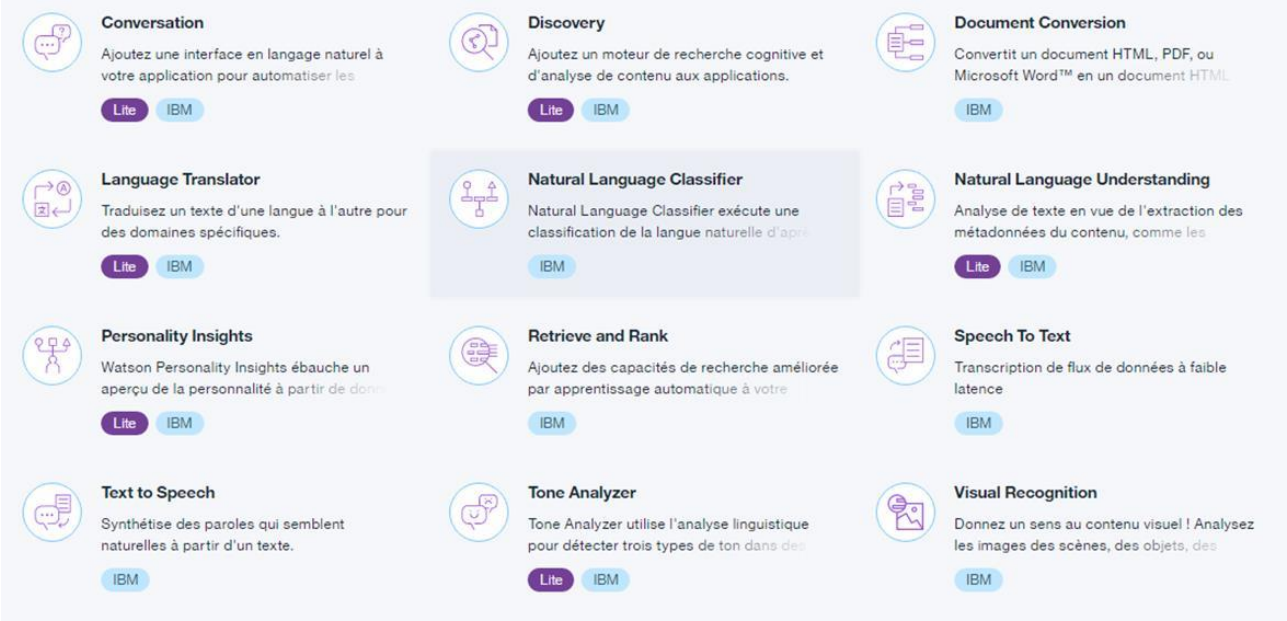
IBM Watson est proposé aux développeurs de solutions sous la forme d'APIs REST[193] qui permettent d'accéder à une large panoplie de services, qui sont intégrées dans la plateforme en cloud Bluemix, avec notamment :

- **Conversation** (anciennement Dialog), un outil qui 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 **limités** car préprogrammés[194].
- **Discovery** (anciennement Retrieve and Rank), un service qui s'appuie sur le logiciel open source Apache Solr et qui permet de traiter les requêtes et questions en s'appuyant sur un mix de moteur de recherche et de machine learning.
- **Visual Recognition** : pour toutes les applications de reconnaissance de l'image, très utilisées en particulier dans le diagnostic à partir d'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.
- **Natural Language Understanding** : qui permet de classifier automatiquement des données textuelles, issues en général de questions posées par des clients en langage naturel.
- **Document Conversion**, un service qui permet de convertir tout document textuel (PDF, Word, HTML) pour les faire ingérer par les services de Watson. C'est l'alimentation de la base de connaissances.

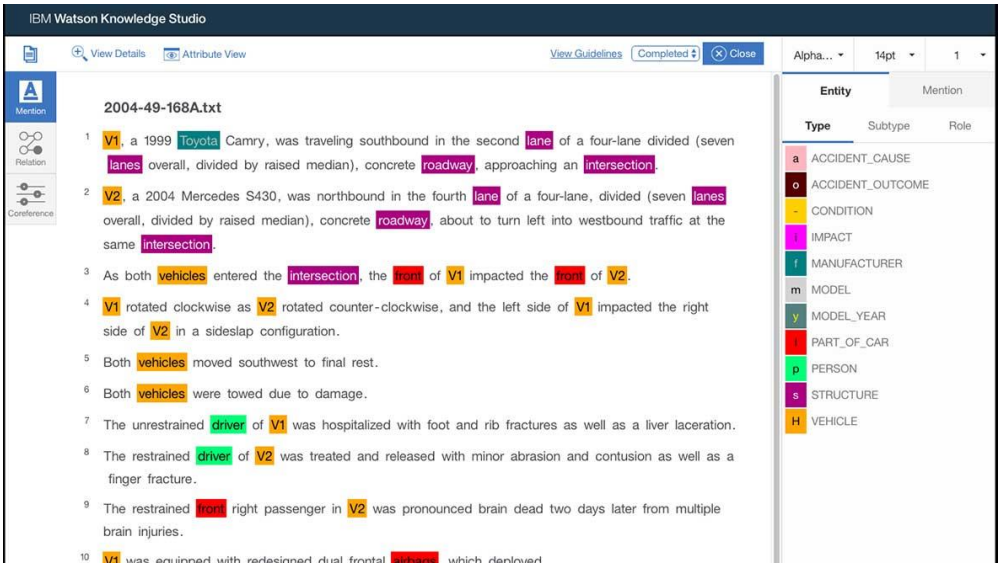
Ces outils sont entraînables via **IBM Watson Knowledge Studio** (*ci-dessous*) qui permet à Watson de digérer le vocabulaire d'un secteur d'activité donnée. C'est de l'apprentissage supervisé qui servira ensuite dans des applications d'analyse de textes ou de création de chatbots.

[193] Avec requêtes http comprenant des GET et des POST et renvoyant le résultat.

[194] 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.



Il faut y ajouter **Watson Explorer** qui permet de créer des solutions personnalisées d'analyse de données structurées et non structurées, c'est une sorte de Netvibes pour Watson.



Ces différentes briques de Watson se retrouvent elles-mêmes intégrées dans des produits tels que **Watson Analytics** pour la compréhension du language naturel.

Watson est aussi décliné sur quelques marchés spécifiques au niveau applicatif avec **Watson Oncology** (cancérologie), **Watson Cybersecurity**, **Watson Virtual Agent** (chatbot)

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[195]. Dans des approches verticales, il faut définir des scénarios de dialogues assez précis

[195] Dans la pratique, Watson sait aussi reconnaitre des images et on peut aussi l'utiliser pour créer un système expert.

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.

L'ensemble est intégré dans les "IBM Cognitive Business Solutions" avec une focalisation sur quelques marchés clés : l'assurance, le retail et la santé, jusqu'à proposer des solutions en apparence clé en main.

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.

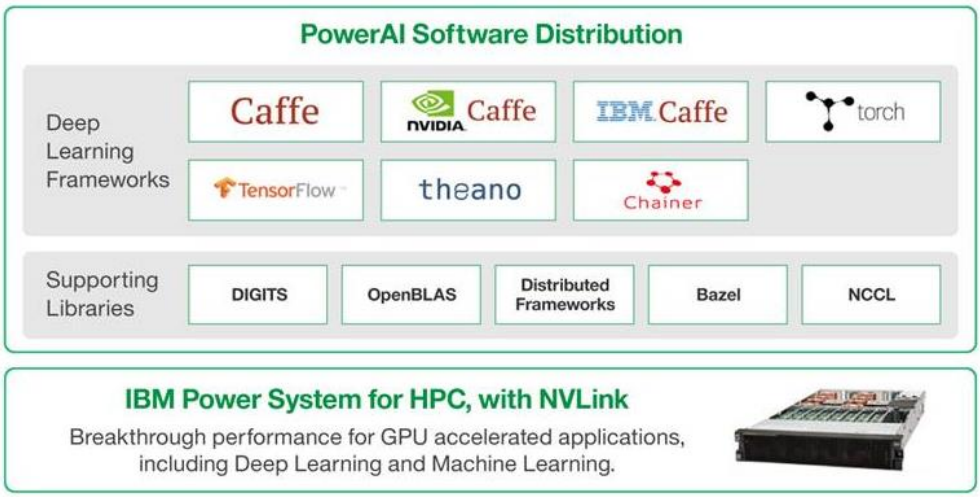
Ces quelques milliers de personnes allouées à Watson sont un bon début mais encore peu au regard des plus de 200 000 collaborateurs d'IBM Services. La migration d'IBM vers un business "cognitif" suffisamment différentié des autres sociétés de services globales dans le monde est une course contre la montre. 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 du prix 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. 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.

Reste à savoir comment se positionne IBM par rapport à l'éventail de solutions du marché. IBM communique peu sur le détail de l'architecture des briques technologiques logicielles que contient Watson. Ou alors, on y trouve des briques logicielles intégrées dans l'offre de manière un peu rapide comme les **Watson Analytics** qui permettent par exemple de segmenter automatiquement une audience client en fonction de ses comportements et d'identifier ceux des segments susceptibles de générer du "churn" (perte de clients).

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 même TensorFlow. Ici, Watson n'est plus spécifiquement de la partie.

Avec leur approche service et intégration, IBM pourra cependant toujours affirmer qu'il sait intégrer les autres briques du marché. Bref, Watson est, en l'état, un objet difficile à benchmarker avec sa concurrence !



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.

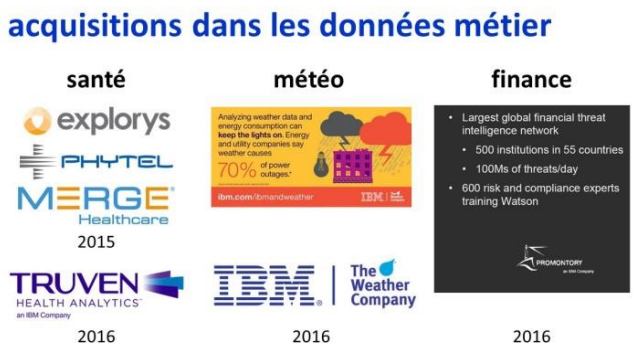
Les projets doivent être longs à closer et à mener avec les grandes entreprises surtout si elles doivent mettre de l'ordre dans leurs données, comme ce fut le cas avec les projets de systèmes experts dans les années 1980. Ils ont probablement également des clients dans les secteurs militaires et du renseignement 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.

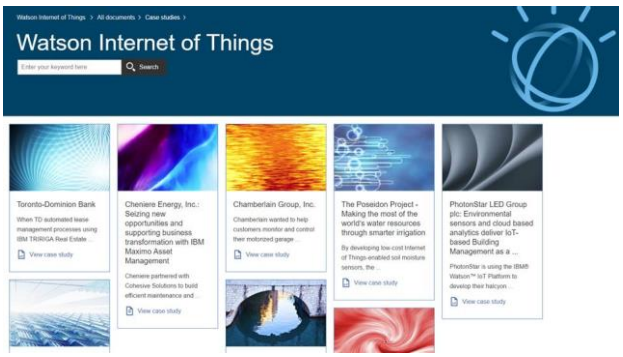
IBM est très dispersé verticalement mais avec un discours assez creux par secteur et relativement peu de références clients hors USA. Les outils marketing et la communication presse d'IBM répète le même discours générique sur Watson avec un zest de

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 quatre acquisitions périphériques aux données de santé : **Explorys** (2009, $15M) avec sa plateforme de cloud dans la santé, **Phytel** (1996, $22,5M) et sa solution de suivi de prise de traitements, **Merge Healhcare** (1987) et ses outils de gestion d'imagerie médicale et **Truven Health Analytics** (2012) et ses outils d'analytics pour les professionnels de santé.





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), créatrice d'un agent conversationnel, **AlchemyAPI** (2005, $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), une startup allemande 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.

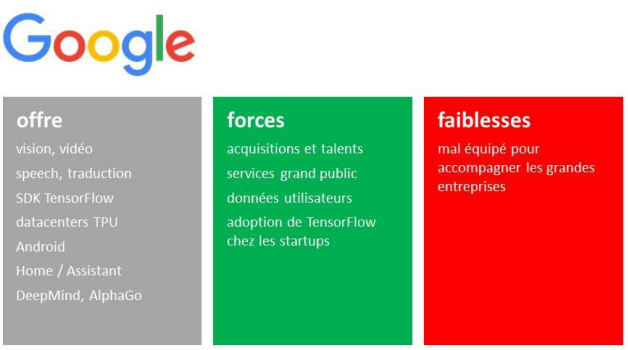
Tout ceci permet à IBM de conserver un peu d'avance dans sa capacité à produire des calculateurs de haute performance. Mais l'industrialisation à grande échelle est le bât qui blesse chez IBM. Pour que ces investissements soient rentables, il leur faudra générer du volume et trouver des débouchés pour ces composants. En effet, pour ce genre de technologie, rien ne dit que l'intégration verticale soit la meilleure approche.

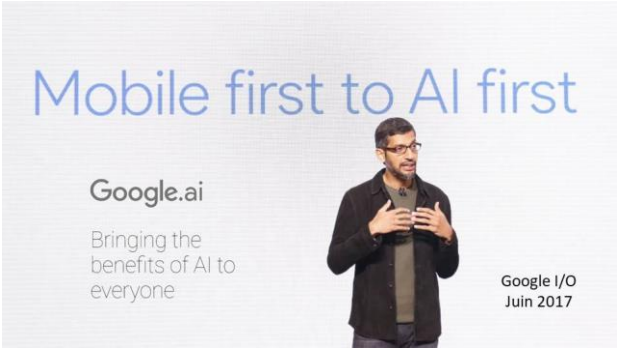
Surtout si la concurrence se structure de manière horizontale comme Intel le fait avec succès sur le marché des PC et des serveurs depuis 35 ans.

Google

Si IBM domine l'actualité de l'IA du côté des entreprises, Google est le roi du pétrole côté services grand public. Et sa communication n'a rien à envier à celle d'IBM. 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[196], dans les agents conversationnels de Google Home et Google Assistant, dans ses Google Car et même dans Android.

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. Mais 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 près de deux milliards d'internautes. Cela lui permet d'accumuler d'énormes volumes d'informations qu'il utilise pour entrainer « 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, ac-

[196] 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.

quis tout ce qui existerait de bien comme compétences dans l'IA. C'est évidemment une vue de l'esprit. 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 limités, 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 l'anglais **DeepMind** 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 d'AlphaGo.

Google avait auparavant mis la main sur la société de reconnaissance vocale **SayNow** en 2011 puis sur **Viewdle** et **PittPatt** 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**, et embauchaient ainsi le canadien Geoff Hinton, considéré comme le père du deep learning.

Ont suivi **Dark Blue Labs** et **Vision Factory**, deux sociétés d'Oxford qui n'ont pas levé de fonds. S'y ajoutèrent la solution de recommandation d'images **JetPack**, le spécialiste de la traduction automatique **Quest Visual**, et celui de la reconnaissance de mouvements **Flutter** 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.

L'année 2014 a 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 Robo-

tics), **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 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).

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 Lee Sedol construite par une équipe d'une vingtaine de personnes de sa filiale DeepMind au jeu de Go contre le champion du monde Sud-Coréen (**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 Deeper 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. Dans "Artificial Intelligence and the Singularity" publié en 2016, Piero Scaruffi se faisait un malin plaisir de relativiser cette victoire (*ci-dessous à droite*) 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 !

La performance a été documentée dans **un article publié dans la revue Nature** en janvier 2016. 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. En mai 2017, une version améliorée d'Alpha Go battait le champion chinois Ke Jie.

Pour la petite histoire, 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.

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 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.

Exhibit 31: Google AI APIs

Google Cloud Platform Machine Learning APIs	Description
Google Cloud Machine Learning Engine - Large Scale Machine Learning Service	Allows you to build sophisticated, large scale machine learning models that cover a broad set of scenarios from building sophisticated regression models to image classification. It is portable, fully managed, and integrated with other Google Cloud Data platform products such as Google Cloud Storage , Google Cloud Dataflow , and Google Cloud Datalab so you can easily train your models.
Google Cloud Jobs API	Provides highly intuitive job search that anticipates what job seekers are looking for and surfaces targeted recommendations that help them discover new opportunities.
Google Cloud Video Intelligence API	Makes videos searchable and discoverable by extracting metadata, identifying key nouns, and annotating the content of the video. By calling an easy-to-use REST API, you can now search every moment of every video file in your catalog and find each occurrence of key nouns as well as its significance. Separate signal from noise, by retrieving relevant information by video, shot, or frame.
Google Cloud Vision API	Enables you to understand the content of an image by encapsulating powerful machine learning models in an easy to use REST API. It quickly classifies images into thousands of categories (e.g. "sailboat", "Eiffel Tower"), detects individual objects and faces within images, and finds and reads printed words contained within images.
Google Cloud Speech API	Enables conversion of audio to text by applying neural network models in an easy to use API. The API recognizes over 80 languages and variants, to support your global user base.
Google Natural Language API	Reveals the structure and meaning of text by offering powerful machine learning models in an easy to use REST API. You can use it to extract information about people, places, events and much more, mentioned in text documents, news articles or blog posts. You can also use it to understand sentiment about your product on social media or parse intent from customer conversations happening in a call center or a messaging app.
Google Cloud Translation API	Provides a simple programmatic interface for translating an arbitrary string into any supported language. Translation API is highly responsive, so websites and applications can integrate with Translation API for fast, dynamic translation of source text from the source language to a target language (e.g. French to English).

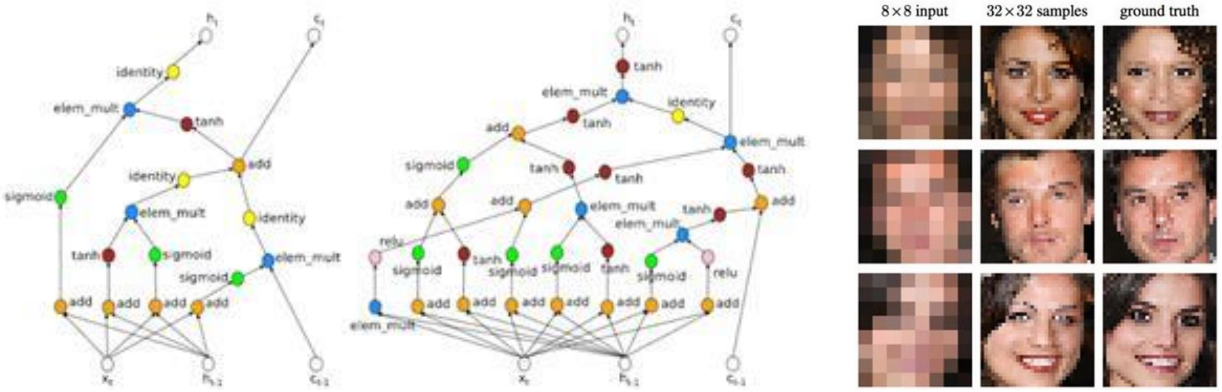
Source: Google, Jefferies

Mais Google regorge de développeurs et d'équipes projets en tout genre. Nous avons par exemple les projets **Google Brain**[197] lancés en 2011 par une petite équipe de chercheurs qui comprend Jeff Dean, Greg Corrado, Andrew Ng et Geoff Hinton depuis 2013.

[197] Google Brain est distinct de Deep Mind qui est basé au Royaume Uni, reste relativement indépendant après son acquisition en 2014.

Cette équipe est à l'origine de systèmes d'encryption évolutifs publiés en octobre 2016[198] et d'un étonnant programme d'amélioration d'images pixellisées publié en en février 2017, exploitant des images de 8x8 pixels[199] 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'entrainement.

En mai 2017, ils publiaient aussi d'étonnant travaux montrant comment ils pouvaient utiliser le machine learning pour améliorer l'arcitecture d'un réseau de neurones qui est d'habitude créée manuellement (*illustrée ci-dessous à gauche*[200]).



L'équipe de Google Brain est aussi à l'origine d'améliorations diverses de Google Translate[201].

Les développeurs de Google sont aussi 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, entrainé sur 260 millions d'images et efficace à 86% en 2016. La méthode ? Une variante de réseau de neurone convolutionnel[202] (*ci-dessus à droite*).

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

---

[198] Qui là aussi a beaucoup fait fantasmer avec « l'IA qui créé son propre langage que les hommes ne comprennent pas ». Cf https://qz.com/822216/google-taught-artificial-intelligence-to-encrypt-messages-on-its-own/.

[199] Cf 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 entrainé avec les images de la dernière colonne.

[200] Et dans Using Machine Learning to Explore Neural Network Architecture, mai 2017.

[201] Cf Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016.

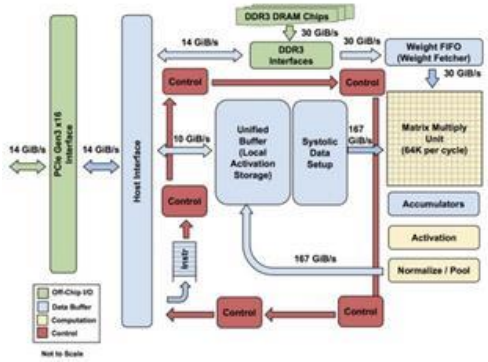
[202] Cf FaceNet: A Unified Embedding for Face Recognition and Clustering, juin 2015. L'un des trois auteurs, James Philbin, a quitté Google en 2015. Il est depuis le directeur de la vision artificielle de **Zoox** (2014, $290M), une startup ultra-bien financée qui veut devenir un opérateur de service de véhicules autonomes.

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 !

Le matériel

Google a développé ses propres processeurs neuromorphiques en 2015/2016, le TPU. Leur architecture est maintenant publique[203]. 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 servent à gagner à AlphaGo et à gérer bien d'autres briques d'AI de Google qui semblent en production.



Mais Google n'envisage pour l'instant pas de commercialiser ces processeurs. Cela devient leur technologie, fabriquée en volume en ASIC, pour équiper leurs 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 se contente de tester avec la NASA les ordinateurs quantiques du canadien **D-Wave**. Rien n'a filtré sur une éventuelle technologie d'ordinateur quantique sortant des laboratoires de Google.

Google est sinon concepteur de produits 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.

Amazon

Vu du monde de l'entreprise, Amazon est un acteur clé de l'IA, essentiellement 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.

[203] Cf https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture/.

Les logiciels

Vu du grand public, il est aussi présent dans l'IA via l'**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 à ce jour plus de 17 500 applications compatibles avec Alexa[204], dénommées « skills » et rien qu'aux USA, dont **Mylestone** (2016, $4M), une application d'une startup de Boston qui se propose de gérer votre mémoire photographique par commande vocale via Alexa.

L'offre d'APIs d'AI pour les développeurs d'applications en cloud comprend, outre Alexa, **Amazon Rekognition**, une fonctionnalité d'analyse d'images 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 reconnaitre. **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 Lex** est le moteur de ges-

[204] Source : https://www.voicebot.ai/amazon-echo-alexa-stats/.

tion de conversations texte et vocale d'Alexa. Et un moteur de traduction doit bientôt voir le jour, issu de la startup américaine **Safaba** (2009) acquise en 2015.

Du côté des couches basses, nous avons **Amazon Machine Learning** (qui 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), **Amazon EMR** (Elastic MapReduce), **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 Solutions for AI		
AI Services	Amazon Rekognition	Image Recognition
	Amazon Polly	Text-to-Speech
	Amazon LEX	Voice and Text Chatbots
AI Platforms	Amazon Machine Learning	
	Amazon EMR	
	Spark & SparkML	
AI Frameworks - AWS Deep Learning AMI	Apache MXNET	
	Tensorflow	
	Caffe	
	Torch	
	Theano	
	CNTK	
	Keras	
AI Infrastructure	Amazon EC2 P2 & G2 GPUS	
	Amazon EC2 P2 & G2 GPUS	
	AWS Lambda	
	Enhanced Networking	
	AWS IoT	
	AWS Greengrass	

Source: Amazon, Jefferies

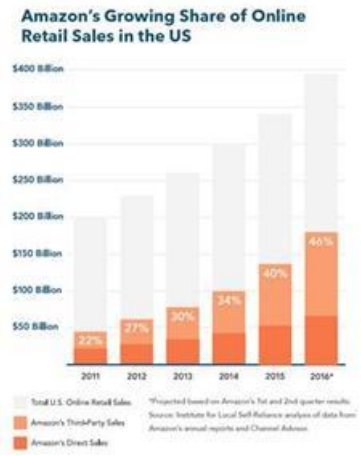
Amazon est comme tous les GAFAMI un acquéreur de startups régulier, mais pas très actif du côté de l'IA. On peut noter celle du spécialiste des robots d'entrepôts **Kiva** (2003, $18M) en 2012 pour $775M, d'**Ivona** (2004), spécialiste du text to speech, acquis en 2013, d'**Angel.ai** (2015, $8M) en septembre 2016, créateur d'un chatbot généraliste qui a certainement du les aider à améliorer Alex et de **harvest.ai** (2014, $2,74M) détection de failles de sécurité, acquis en 2017.

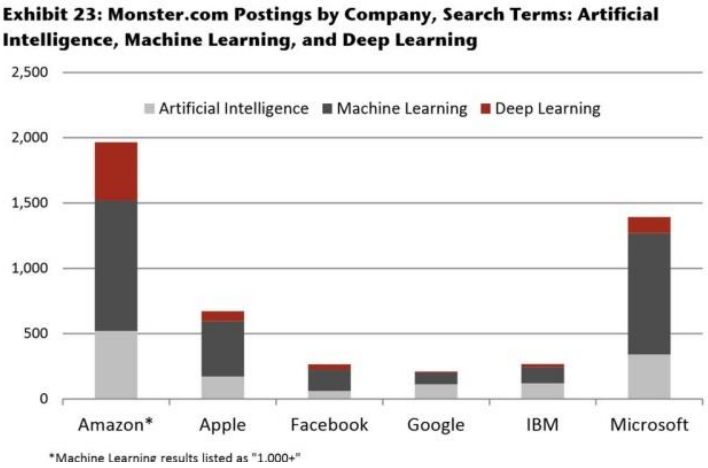
Les données

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 (*cf schéma ci-dessous à gauche*). 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.

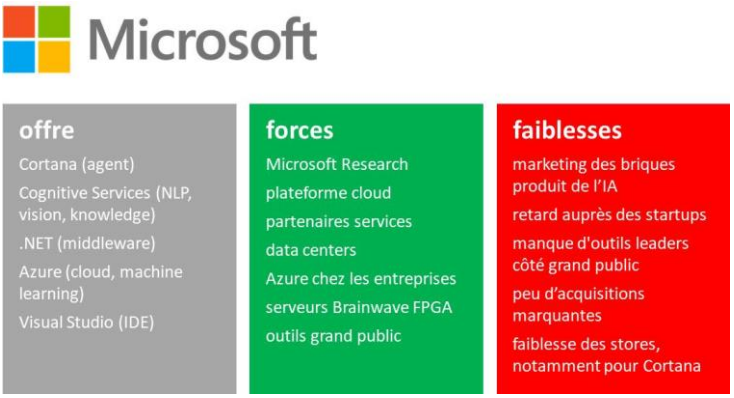
Cela explique probablement pourquoi Amazon est l'un des plus gros recruteurs aux USA de spécialistes de machine learning[205] (*chart ci-dessous à droite*).





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. De manière assez classique, son offre couvre le traitement du langage, de la vision artificielle, des données et la gestion des connaissances. Mais 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.



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'intelligence artificielle.

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.

[205] Source : IBM (IBM) Creating Shareholder Value with AI? Not so Elementary, My Dear Watson de Jefferies, juillet 2017.

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 ! Chez Google, la frontière entre recherche et développement est plus floue.

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 business.

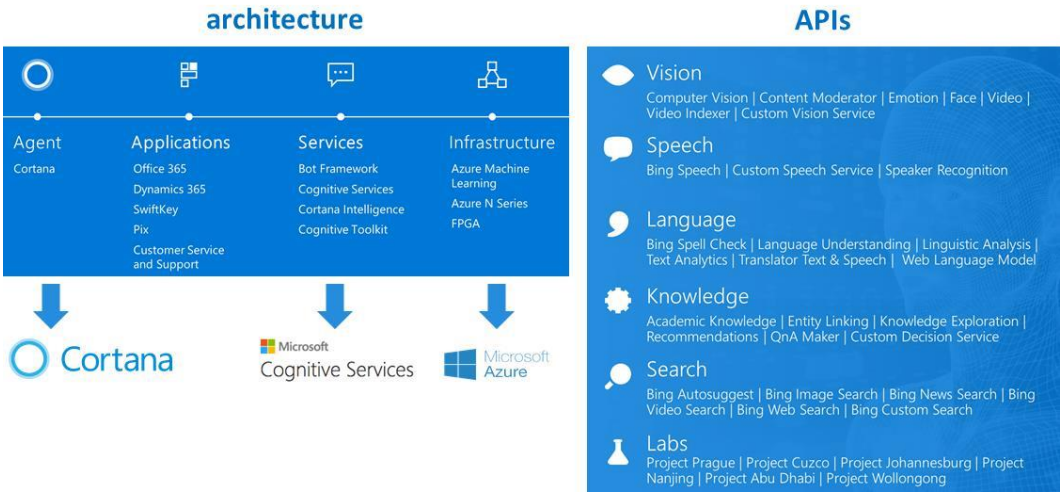
Microsoft qui est maintenant résolument tourné vers le cloud fait tout de même quelques acquisitions de startups pour accélérer son "time to market" dans l'IA ou dans la périphérie de l'IA. Les équipes de recherche fondamentale travaillent en effet sur des domaines où le risque est plus scientifique 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 peu nombreuses chez Microsoft. On peut citer **Revolution Analytics**, 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**, un logiciel de clavier virtuel mobile qui s'appuierait lui aussi sur du machine learning. En 2015, Microsoft avait aussi mis la main sur **Prismatic**, un agrégateur de news s'appuyant sur du machine learning, ainsi que **Double Labs**, une application Android de notification elle aussi basée sur du machine learning. En 2017, c'était au tour de **Genee** (2014, $1,45M) un gestionnaire d'agenda virtuel et du Canadien **Maluuba** (2011, $8,2M), un spécialiste du deep

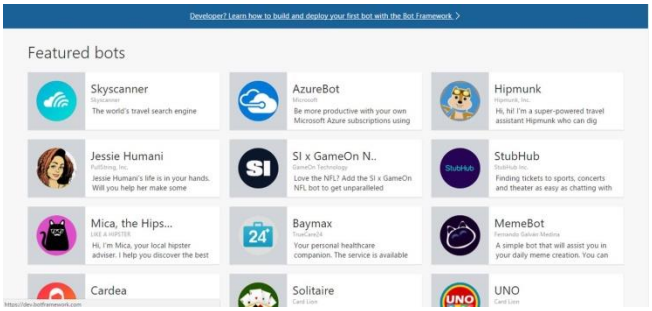
learning appliqué au traitement du langage qui planche sur l'AGI (Artificial General Intelligence), dont les équipes se sont faites remarquer en faisant gagner leur solution au Pac-Man avec leur technique Hybrid Reward Architecture[206] (vidéo).

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, 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**).



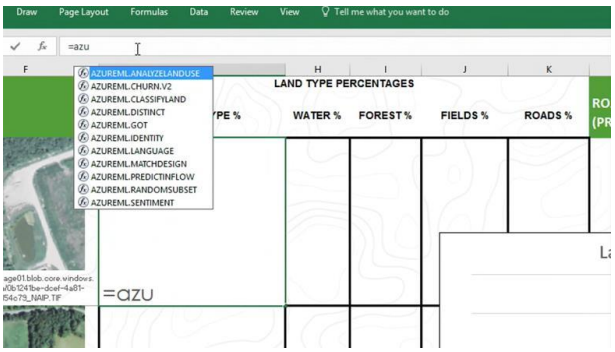
L'architecture d'IA de Microsoft s'appuie sur :

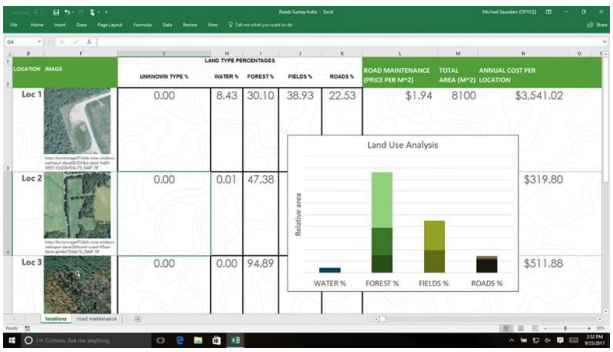
- L'**agent Cortana**, bon à tout faire, qui répond à la voix, joue le rôle de traducteur. C'est un peu l'équivalent d'Amazon Alexa et de Google Assistant. Quelques bots ont été développés avec les APIs de Microsoft mais l'offre n'a rien à voir avec l'abondance autour d'Amazon Alexa.



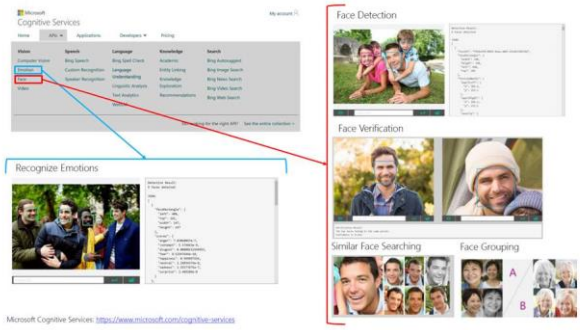
- Diverses **applications** qui intègrent des briques d'IA, comme Office 365, Dynamics 365 et l'application mobile de gestion de photos Pix. Le traitement du langage à base d'IA est disséminé dans Office, et depuis des années. En septembre 2017, Microsoft annonçait l'intégration de services de machine learning du cloud Azure dans Excel, qui se manifestent sous la forme de fonctions (*ci-dessous*).

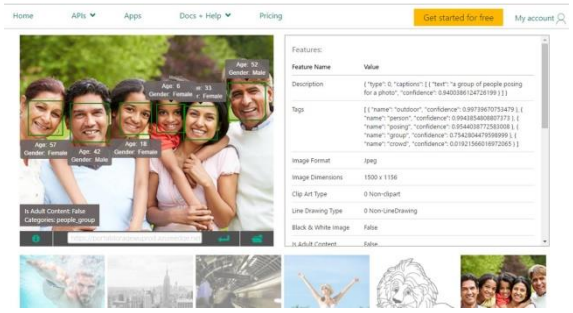
[206] Cf Hybrid Reward Architecture for Reinforcement Learning, juin 2017, qui décrit une architecture d'apprentissage par renforcement avec des agents fonctionnant en parallèle.



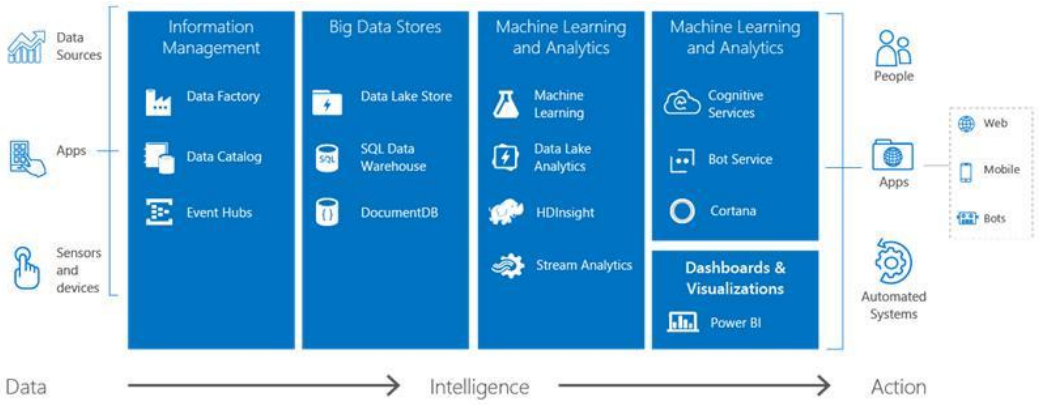


- Les **services cognitifs** qui comprennent plus d'une vingtaine d'APIs sensorielles qui font de la reconnaissance d'images, du traitement du langage naturel (NLP), de la gestion de connaissances et de la recherche. A bas niveau, Microsoft propose en open source son framework de "deep learning" **CNTK** (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*).



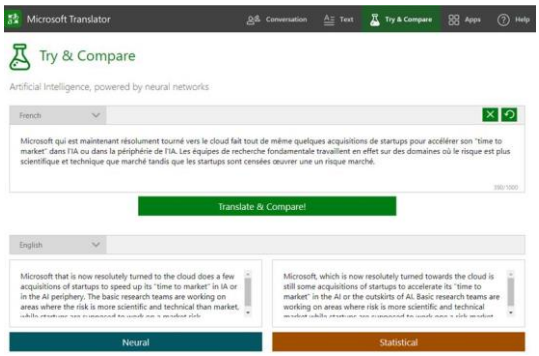


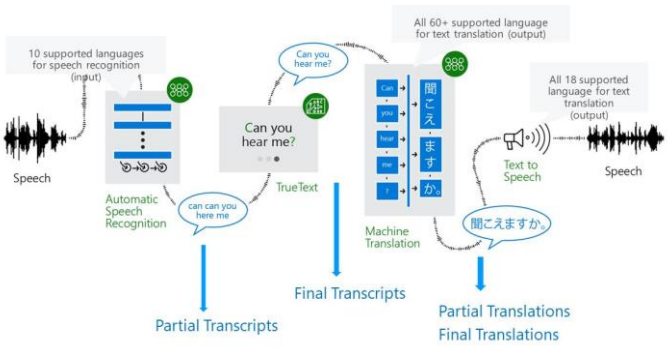
- Des **outils de création** d'applications avec l'IDE Visual Studio (Integrated Development Environment) et ses Code Tools for AI, et aussi l'outil Azure Machine Learning Studio, lancé en 2015, qui permet de créer ses modèles de machine learning et de les mettre en production. Annoncé en septembre 2017, 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).



- L'infrastructure en cloud **Azure** et les logiciels serveurs de Microsoft.
- Du **matériel spécifique**, notamment des processeurs neuromorphiques développés dans des chipsets FPGA.

Dans ces différents étages, 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 données

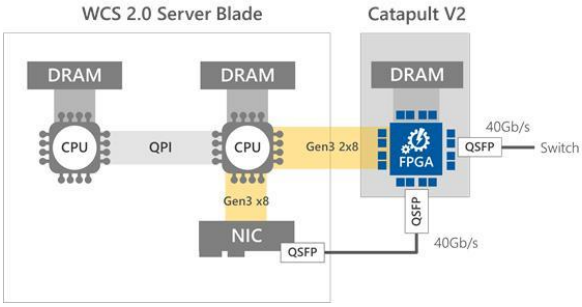
Microsoft dispose d'une activité grand public, certes pas aussi soutenue que celle de Google, mais qui lui permet d'avoir une forte expertise dans le cloud ainsi que dans la captation de données d'usages, à même d'alimenter ses outils de machine et deep learning. Il en va ainsi du moteur de recherche **Bing**, de **Skype**, de **MSN**, de la console de jeu **Xbox** et bien évidemment de **Windows**.

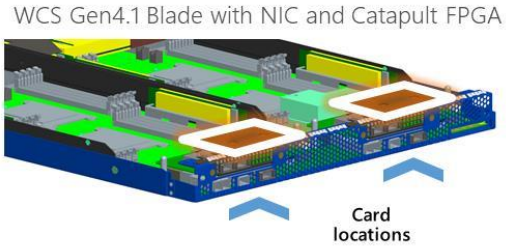
Le matériel

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[207] et qui s'appuient sur :

[207] Cf https://www.microsoft.com/en-us/research/blog/microsoft-unveils-project-brainwave/.

- 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)[208]. 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 qui plus est optimisée à la fois pour des réseaux de neurones convolutionnels (CNN, pour le traitement de l'image) qui nécessitent de multiplier des matrices et des réseaux de neurones récurrents (RNN, pour le traitement de la parole et 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 les ordinateurs quantiques mais il est encore trop tôt pour en parler d'un point de vue pratique.

Oracle

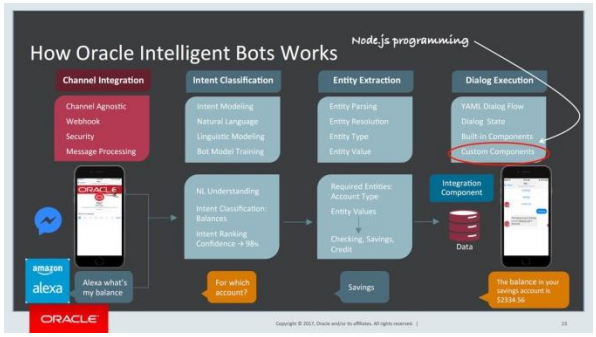
Avec SAP, Oracle est l'un des plus gros éditeurs de logiciels d'entreprise au monde. Il est plutôt discret dans son intégration de l'IA dans la stratégie. Comme l'éditeur gère de gros volumes de données depuis des décennies, ils doivent exploiter des techniques s'apparentant au machine learning depuis quelques temps, sans forcément le marketer.

[208] 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).

Ils se sont mis à le faire en intégrant progressivement et ouvertement divers outils de machine learning dans leurs logiciels d'infrastructures horizontaux et applications métiers verticales.

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 aussi 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.



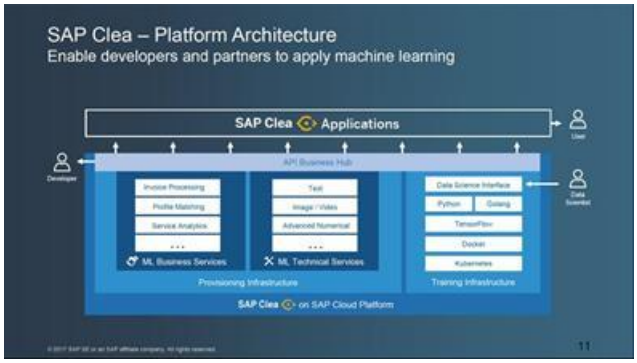
Quid des acquisitions dans l'IA ? Oracle est assez friand de startups, mais curieusement, pas vraiment dans l'IA. On note surtout celle de l'israélien **Crosswire** (2013, $5M) en 2016 qui propose une solution cross-device de ciblage et d'analytics publicitaires doté d'outils de présentation graphique.

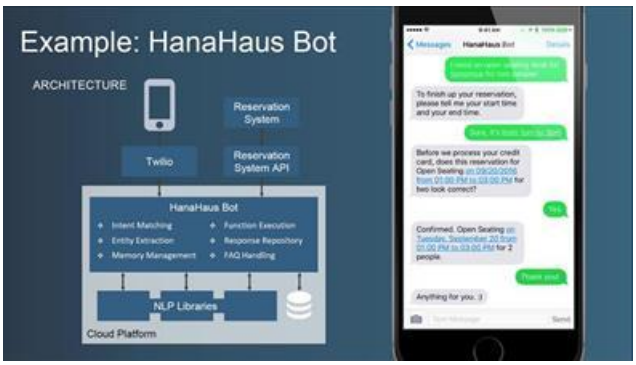
SAP

SAP a adopté vis à vis de l'IA une position voisine de celle d'Oracle, assez discrète, et intégrant progressivement l'IA dans son offre.

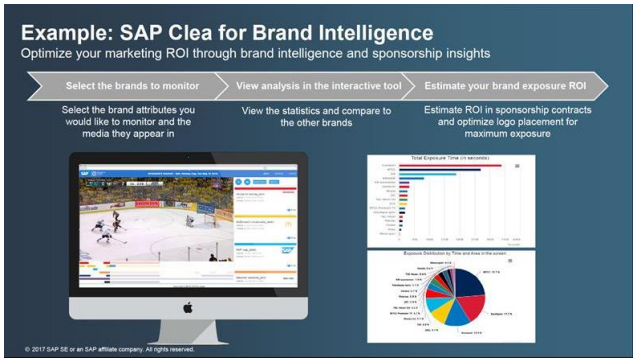
Nous avons dans l'ordre, de bas en haut :

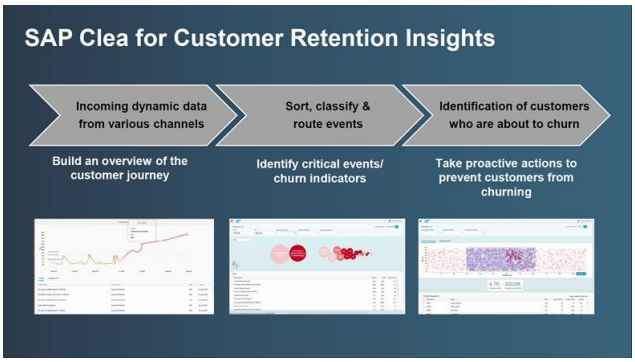
- **SAP Clea** qui comprend des services de machine learning et de deep learning pour les développeurs avec traitement classiques langage, vision et données. La nature de ces briques technologiques ni leur origine n'est précisée dans la littérature marketing de SAP. Curieuse pratique. En même temps, SAP n'est pas le roi du pétrole dans le recrutement de développeurs et surtout de ceux des startups.



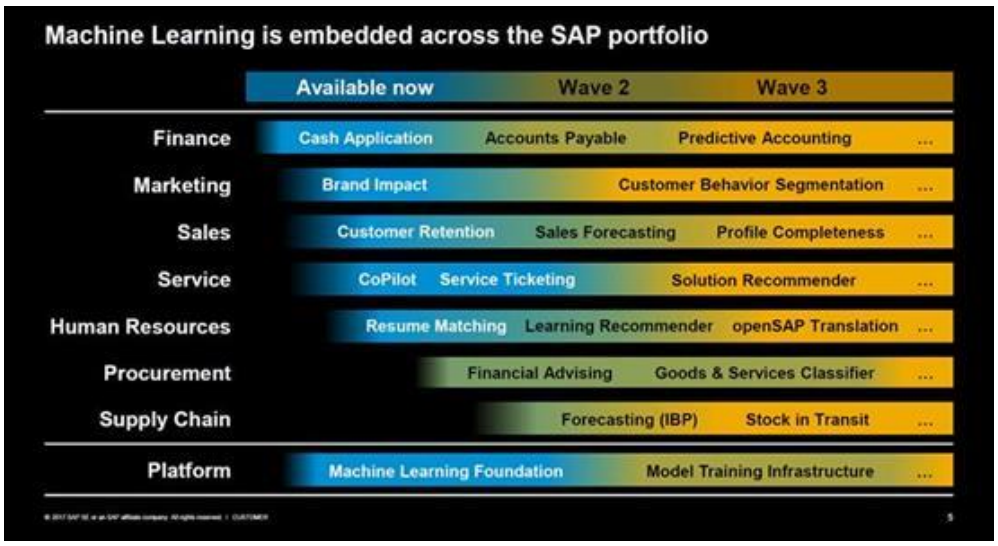


- **Leonardo** qui fait partie de la plateforme HANA, avec divers outils de machine learning non précisés.





- L'intégration de ces technologies dans diverses solutions applicatives comme dans les outils d'analytics issus de l'acquisition en 2008 de Business Objects, les SAP BusinessObjects Cloud for Analytics. Elle se fait étape par étape comme l'indique la roadmap ci-dessous.



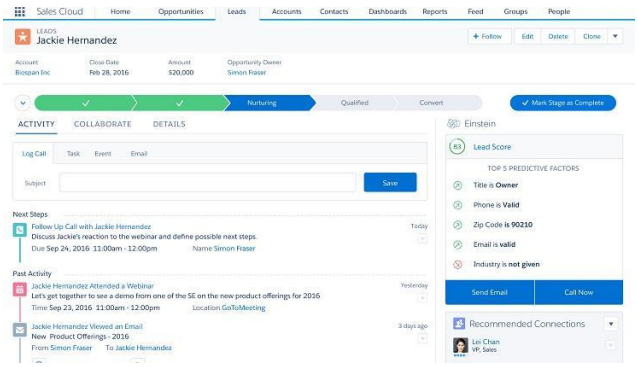
Salesforce

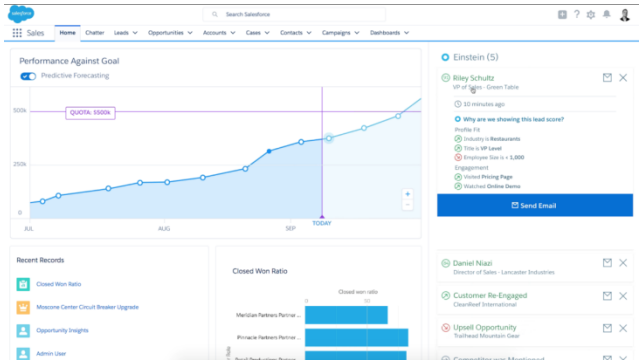
Chez Salesforce, l'offre d'IA s'appelle modestement **Einstein**, une offre d'IA en cloud au service des forces de vente.

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.





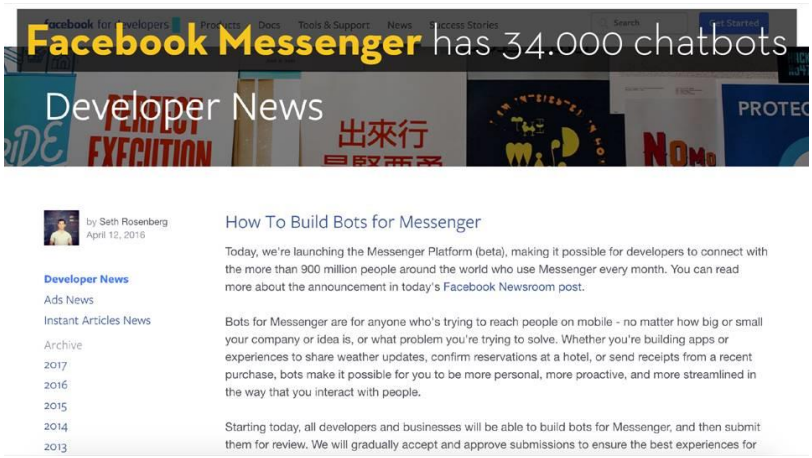
Une bonne partie d'Einstein provient de briques et de compétences récupérées par SalesForce à l'occasion de l'acquisition de diverses startups : **RelateIQ** (2011, $69M) en 2014, qui était spécialisées dans la relation client, devenu SalesForceIQ, une version de SalesForce pour les PME, **PredictionIO** (2013, $2,65M) acquis en 2016 pour sa solution open source de machine learning et **MetaMind** (2014, $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**.

Facebook

Le leader mondial des réseaux sociaux est avide d'IA à tous les étages pour améliorer l'ensemble de ses services, qui 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** servant à la création de chatbots s'intégrant dans l'application Facebook Messenger et lancé en avril 2016.

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 le FAIR (Facebook Artifical Intelligence Research) à Paris depuis 2015. Yann LeCun est aussi depuis peu professeur au Collège de France sur le deep learning[209].

[209] Voir sa **leçon inaugurale** qui fait un très bon panorama technique du machine learning.



Parmis les projets de recherche, **DeepFace** est une solution de deep learning de reconnaissance des visages avec un réseau de neurones de neufs couches et 120 million connexions entrainé sur quatre millions d'images. La precision du système serait de 97%. Le système a bénéficié de la contribution de Yaniv Taigman, issu de **Face.com**, acquise en 2007. Facebook veut utiliser la reconnaissance d'images pour informer ses utilisateurs mal-voyants du contenu visuel de leur timeline.

Les autres acquisitions de Facebook dans l'IA comprennent :

- **Jibbigo** (2009) en 2013, pour son système de traduction speech-to-speech.
- **Pebbles Interfaces** (2010, $14,45M) en 2015, pour son système de captation de gestes.
- **wit.ai** (2013, $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.
- **Surreal Vision** (2014), une startup anglaise acquise par Oculus en 2015, spécialisée dans vision 3D.
- **Faciometrics** (2015) en 2016, pour sa solution mobile d'analyse de visages.
- **Ozlo** (2014, $14M) fin juillet 2017, qui aide à trouver un bon restaurant.

Les équipes de Facebook planchent aussi sur la reconnaissance automatique de sports dans les vidéos ou de chiens dans les photos. Facebook a aussi développé une fonction qui décrit le contenu de photos, adaptée aux aveugles, presque simultanément à une fonction du même genre proposée aux aveugles par Microsoft.

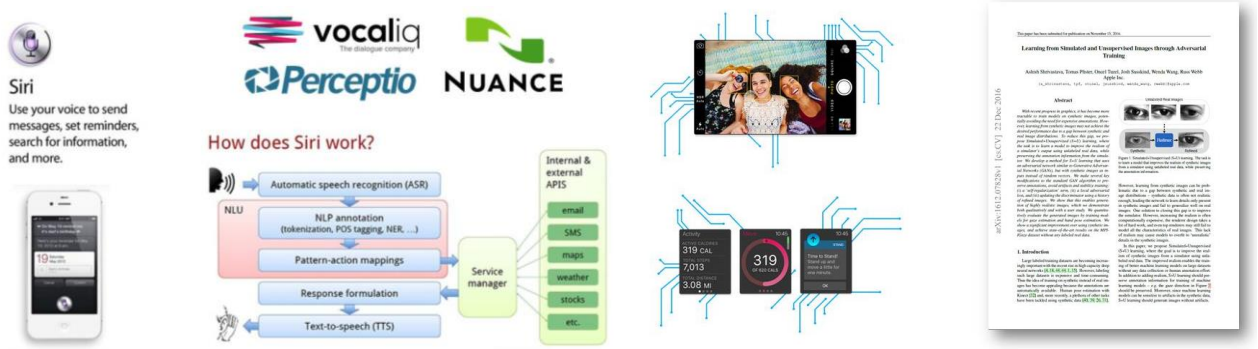
Le géant des réseaux sociaux rêve aussi probablement de créer des solutions de marketing ultra-intelligentes, capables de devenir les aspirations et intentions des utilisateurs. Par exemple, une solution qui saura que je change de ville tous les ans pour mes vacances et évitera de m'exposer à des publicités liées à des villes déjà visitées !

Enfin, on ne peut pas négliger les applications potentielles de l'IA dans la réalité augmentée. C'est un enjeu pour Facebook (Oculus Rift), pour Google (qui finance Magic Leap dans ce domaine) et Apple (qui aurait un projet dans le domaine).

Prochain gros projet : détecter les fake news ! Vaste programme !

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 a cependant publié un premier papier fin 2016 sur la reconnaissance d'images[210].



Les acquisitions d'Apple sont peu nombreuses en règle général. Dans l'IA, on peut compter **Turi** (2013, $25M) en 2016 pour $200M qui avait développé un projet d'analytics à base de machine learning, **Emotient** (2006, $6M) en 2016 pour la reconnaissance des visages et des émotions, **VocalIQ** (2011, $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) en 2015, dans la reconnaissance d'images s'appuyant sur du deep learning. SIRI est de son côté le résultat de l'acquisition en 2010 de la startup **SIRI** (2007, $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 ! 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 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.

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, la grande bête noire de Steve Jobs.

---

[210] Cf 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 automatique.

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 vu la forte dépendance entre les solutions d'IA et les données qui les alimentent.

Caractéristiques

Pour ce qui est des startups, l'IA est devenue un phénomène de mode comme l'ont été les réseaux sociaux (vers 2004), la vidéo (vers 2006), la mobilité (à partir de 2009), les objets connectés (vers 2011) ou le cloud. 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 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[211]. Ces évaluations ont toujours tendance à gonfler des chiffres qui sont dans la pratique incalculable. 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[212].

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.

Depuis début 2016, ma position sur la question a cependant 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 boites à outils prêtes à l'emploi comme chez Clarifait pour l'analyse d'images.

---

[211] Cf Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $12.5 Billion This Year, According to New IDC Spending Guide, avril 2017.

[212] Que j'avais eu l'occasion de décrire dans La grande intox des objets connectés en août 2015.

L'IA est en train de devenir l'équivalent moderne du développement web : un ensemble de techniques de plus en plus abordables. 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 !

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. 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'entrainement 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[213].

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 :

- **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.
- **Métier** : une bonne startup connait bien le métier de ses clients, ses contraintes, ses besoins et aspirations. Elle sait créer une solution qui s'intègre bien dans l'existant.

[213] Source : IDC FutureScape: Worldwide IT Industry 2017 Predictions.

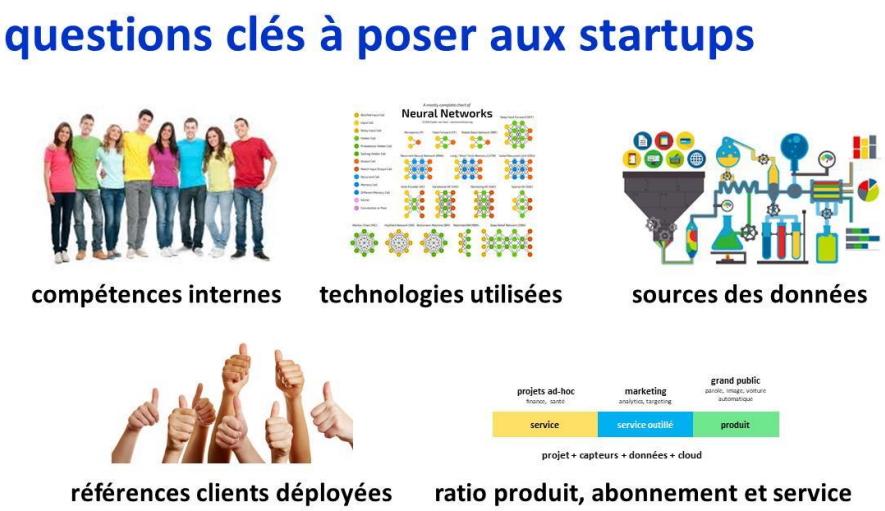
- **Données** : il n'y a pas d'IA sans accès à des données pour entrainer ses modèles. Les données exploitées par la startup peuvent être de trois types : ouvertes et facilement disponibles sur Internet (open data, ImageNet[214], WordNet, MNIST), collectées de manière exclusive à la startup, par exemple 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 prête à respecter la nouvelle règlementation européenne RGDP en mai 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 ? 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 ?
- **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.

- **Business** : quel est le modèle économique de la startup ? Est-il récurrent ? Où sont les économies d'échelle ? Des questions sommes toutes 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.

[214] 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.

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'artisanal. Cela n'apparait 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".

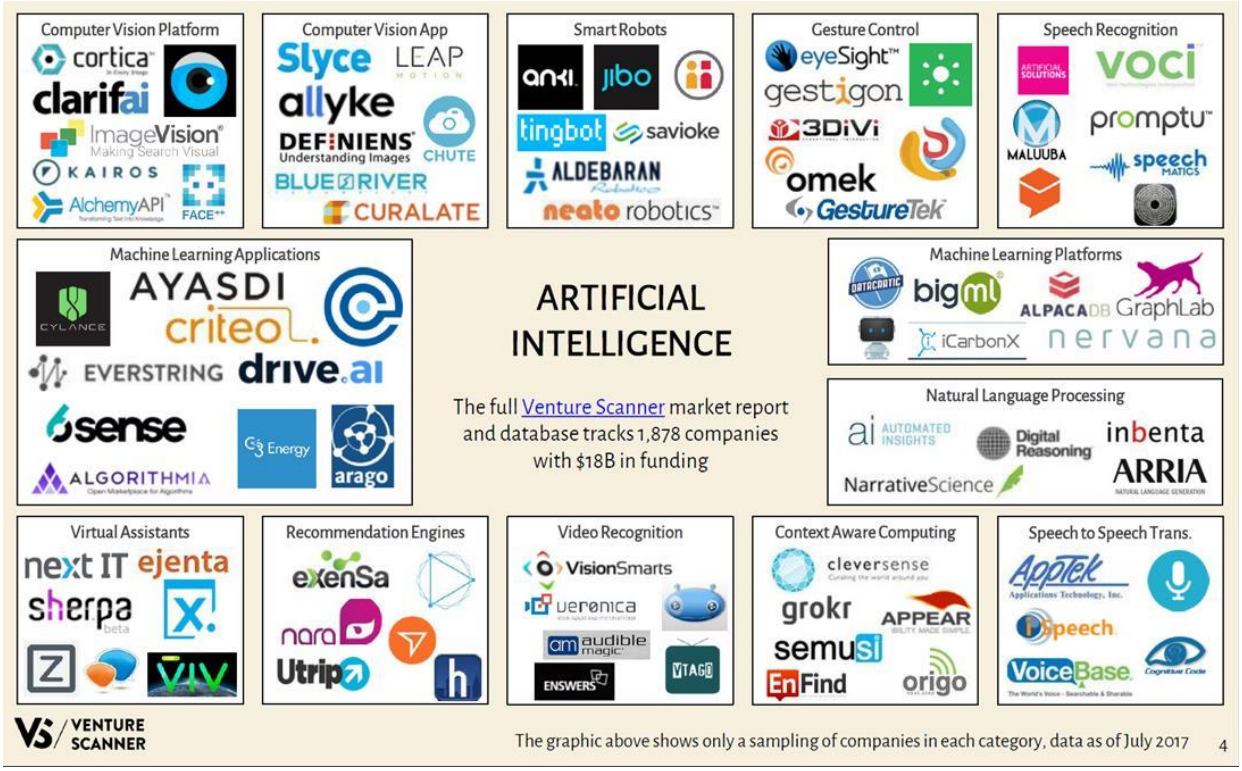


Et aussi, ne pas oublier d'avoir une démonstration du logiciel ! Dans l'IA, l'ergonomie est aussi importante que la fonction !

Cartographie

Je vais m'appuyer sur ce suivi du secteur par le site **VentureScanner** qui était actualisé en juillet 2017[215]. Il organise le marché des startups de l'intelligence artificielle en 13 segments et évalue leur ancienneté et leur financement.

[215] Cf https://www.venturescanner.com/blog/2017/artificial-intelligence-market-overview-and-innovation-quadrant-q3-2017. En juillet 2017, ils suivaient 1888 startups dans l'IA dans 13 catégories sur 70 pays représentant $19B de levées de fonds.

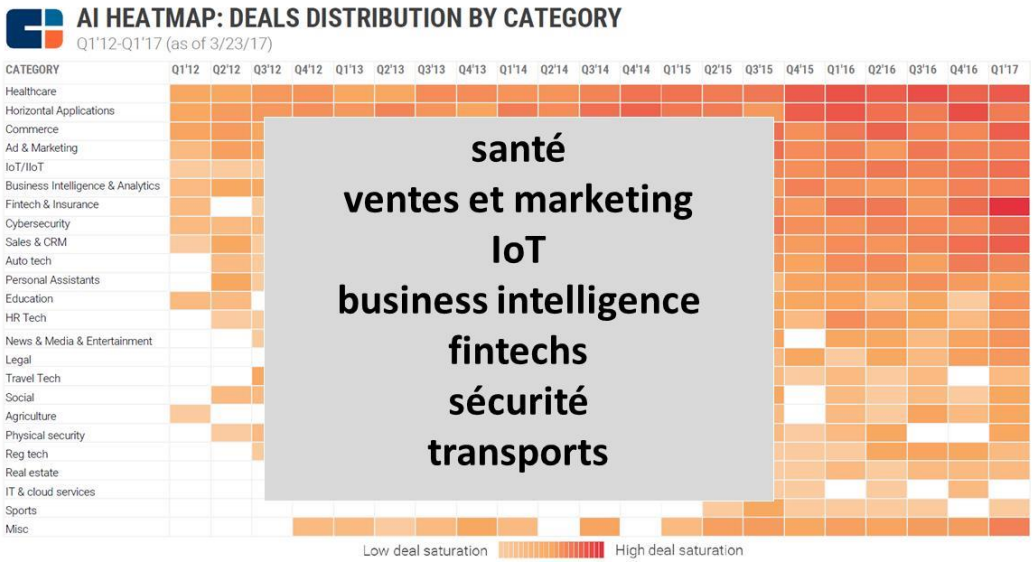


Et voici leur découpage, un peu restructuré pour le simplifier :

- **Plateformes** : machine learning, deep learning, réseaux de neurones, composants
    - **Plateformes de deep learning et machine learning** (256 startups, $2,9B) : qui font avancer l'état de l'art côté algorithmie, avec des modèles prédictifs divers.
- **Données** : prédictif, analytics, recommandation, en gros, ce qui ne concerne pas le cognitif, la vision et le langage
    - **Systèmes de recommandation** (102 startups, $1,5B) : pour prédire les comportements des utilisateurs. On en trouve notamment dans la musique, la vidéo et la restauration.
- **Vision** : recherche d'images, commande gestuelle
    - **Applications vision** (214 startups, $1,1B) : appliquée à des marchés verticaux comme dans le ecommerce ou la santé.
    - **Plateformes vision** (191 startups, $2B) : technologies horizontales dans la vision artificielle comme les moteurs de recherche d'images ou les systèmes de tagging d'images.
    - **Reconnaissance de vidéos** (24 startups, $104M) : comme pour détecter les contenus protégés.
    - **Commande gestuelle** (60 startups, $694M) : à la frontière entre objets connectés et captation de mouvements et d'images.
- **Langage** : chatbots, traduction, extraction, recherche

- **Traitement du langage** (304 startups, $2,2B) : solutions techniques dans le domaine pour comprendre le langage, le traduire, le résumer, le générer, etc.
- **Reconnaissance de la parole** (163 startups, $1,1B) : avec des logiciels de reconnaissance de la parole, fournis en cloud ou en mode embarqué.
- **Traduction vocale** (21 startups, $45M) : traduction « speech to speech » utilisable dans des contextes divers comme dans les chats vidéos.
- **Assistants virtuels** (186 startups, $1,3B) : les fameux chatbots, qui sont déclinés en plateformes de chatbots et en chatbot pour marchés verticaux ou horizontaux, en b2b et b2c.

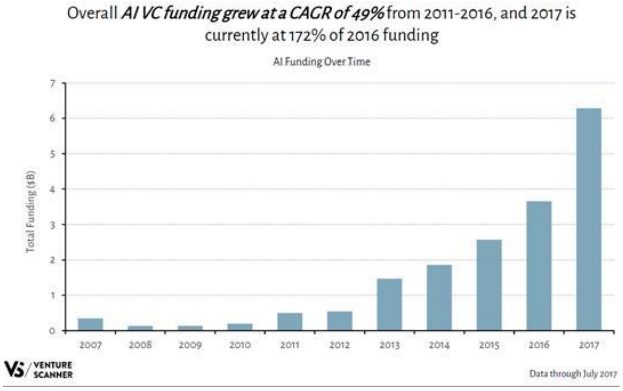
- **Robots** : plateformes, humanoides, services
  - **Robots** (168 startups, $2,8B) : du robot domestique au robot industriel.
- **Métier** : avec applications verticales (transports, santé, finance, ecommerce, juridique, agriculture) et horizontales (sécurité informatique, RH, marketing)
  - **Applications métier** (570 startups, $8,3B) : 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. Un schéma de CBInsights 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, $132M) : 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.

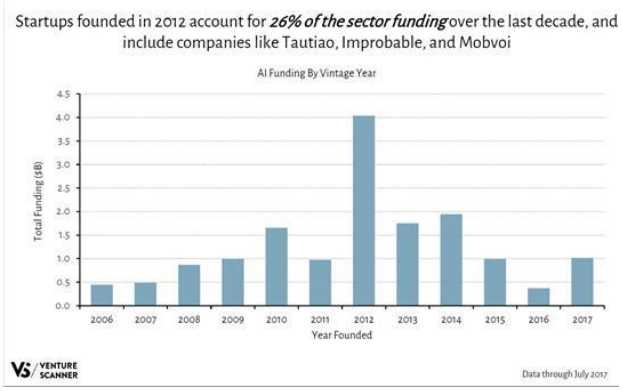


Dans la pratique, la frontière entre plateformes génériques et applications métiers est ténue. Nous avons par exemple vu dans le domaine de la santé des startups spécialisées dans l'imagerie médicale sur une seule pathologie (rétinopathie diabétique, dermatologie) ou d'autres qui en couvraient plusieurs. 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 2017 (cf les schémas ci-dessous de Venture Scanner de juillet 2017).





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.

Caractéristiques

Au-delà des facteurs clés de succès évoqués précédemment, 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.
- 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 !
- On retrouve aussi beaucoup d'anciens de l'université de **Stanford** et du **MIT** dans les startups de l'IA, généralement bardés d'un ou de plusieurs PhD en IA.
- 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 ef-

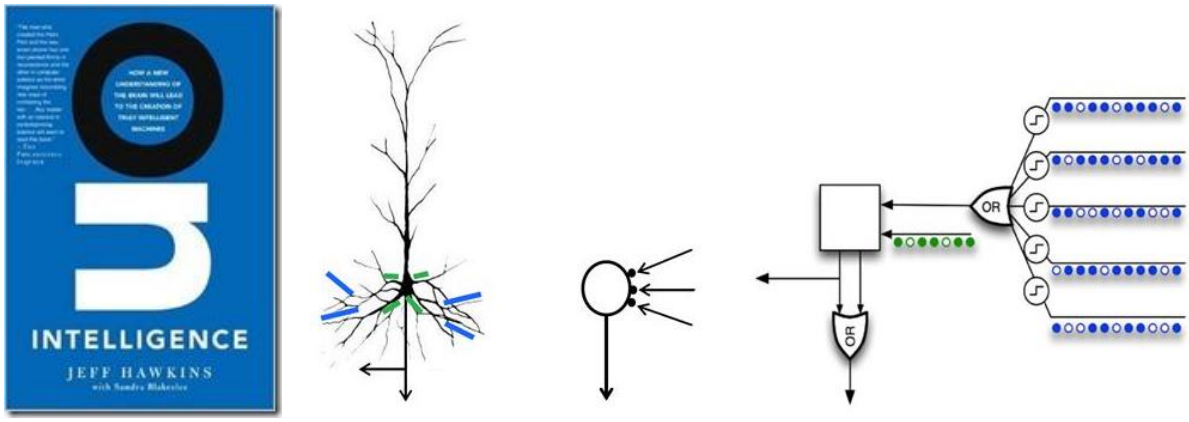
fet push/pull courant dans le grand public (la demande pour des smartphones Android entrainant celles d'applications tournant dessus).

- 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 presque toutes des startups grand public.

Plateformes de deep learning

C'est une catégorie de startups importante en volume mais aussi la plus déroutante car difficile à évaluer. Voici un tour d'horizon de quelques-uns de ses acteurs, notamment les plus visibles d'entre eux.

Numenta (2005) est une société lancée par le créateur de Palm, Jeff Hawkins. 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[216]. 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.

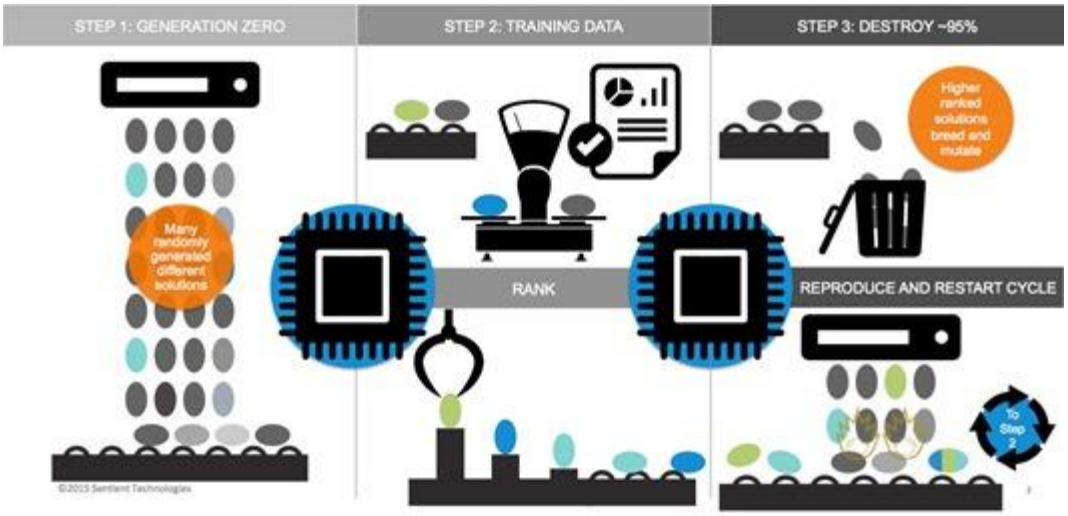
[216] L'ouvrage est téléchargeable gratuitement ici : https://papers.harvie.cz/unsorted/Jeff%20Hawkins%20-%20On%20Intelligence.pdf.

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[217].

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.

Sentient Technologies (2007, $143M) développe pour sa part 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.



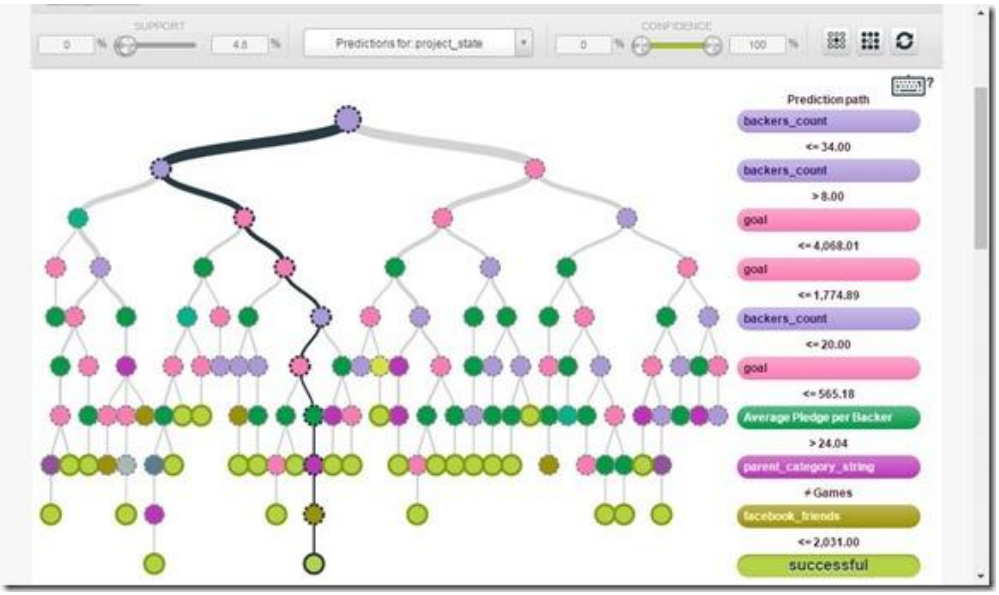
Digital Reasoning (2000, $73M) a été créée par des anciens d'Oracle et de la CIA (entre autres provenances) et est financée par In-Q-Tel, le fonds d'investissement de cette dernière. Sa solution d'analyse de données est utilisée par le renseignement et la défense US ainsi que dans la finance. Comme celle 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).

[217] Voir ces critiques chez **Jeff Kramer**, **Ben Goertzel** et sur **Quora**.

Scaled Inference (2014, $8M) propose une plateforme de machine learning en cloud via des APIs. Elle comprend de la reconnaissance de formes, des détectons d'anomalies, des algorithmes de prévision. Startup créé par un ancien de Google. Solution pas encore disponible.

Skymind (2014, $3,3M) 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.

BigMl (2011, $1,63M) a l'air d'être un outil d'analyse assez générique qui analyse les comportements clients, permet du diagnostic de matériel, dans la santé, dans les risques pour des prêts. L'ensemble s'utilise via des APIs attaquant un service en cloud. Au moins, leur site fournit des **exemples** de traitement de jeux de données comme ce modèle prédictif de succès de campagne de financement participatif sur Kickstarter en fonction de leurs différentes caractéristiques. Intéressant !



Cycorp (1994) est une sorte de laboratoire de recherche privé en IA financé par des contrats du gouvernement US, dont la DARPA, et d'entreprises privées. Le projet de recherche Cyc dont il est issu a plus de 30 ans au compteur ! Il vise à modéliser les connaissances et à permettre d'automatiser la recherche scientifique. 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.

Ayadsi (2008, $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.

Narrative Science (2010, $40M) propose Quill, une plateforme qui analyse les données structurées et non structurées issues de sources diverses pour en extraire ce qui est important et en produire des résumés automatiquement. La solution permet notamment d'exploiter les données issues de Google Analytics ou d'historique de transactions financières (*ci-dessous*). Startup créée par un ancien de Google et de Carnegie Mellon.

Synapsify (2012, $1,45M) a créé CORE, un outil d'analyse et de traitement en langage naturel qui fait de la recommandation de contenus.

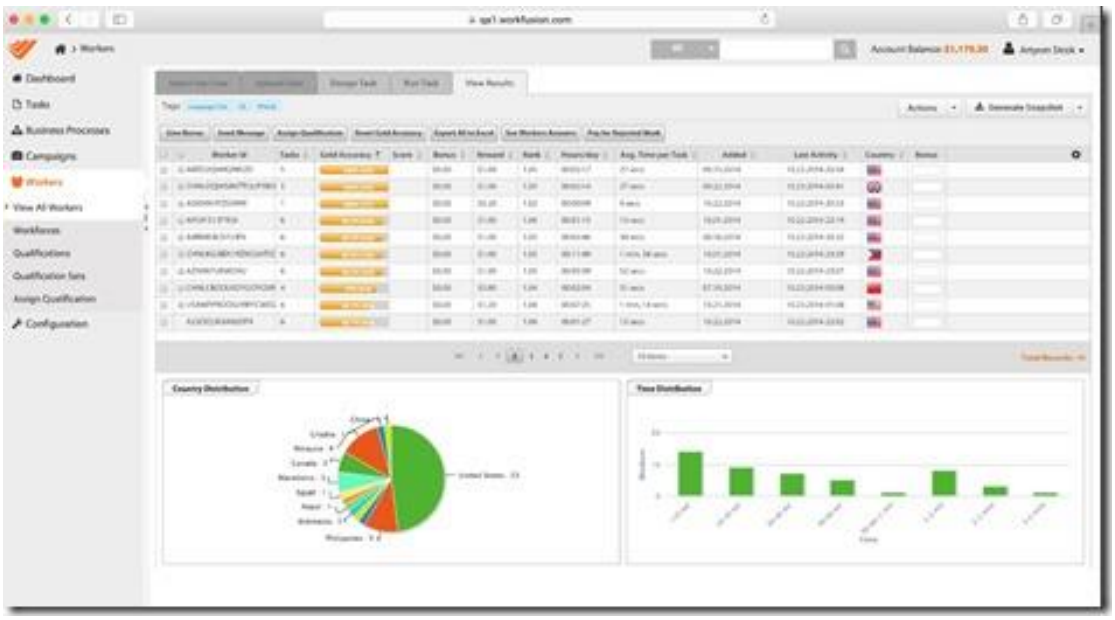
Idibon (2012, $6,9M) analyse les textes structurés, notamment issus des réseaux sociaux, pour les classifier automatiquement et réaliser des analyses statistiques dessus.

Moteurs d'analyses prédictives

Les startups de ce domaine proposent des outils d'ingestion et d'analyse de gros volumes de données structurées et non structurées (documents, images, etc). Les outils d'analyse s'appuient sur un panaché de méthodes associant des statistiques, du data mining, du machine learning et du deep learning). Certains proposent leur solution en open source et la plupart les diffusent surtout en cloud.

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.

Work Fusion (2010, $71,3m) 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.



Sentenai (2015, $1,8m) propose une plateforme d'analyse prédictive, en cloud, qui est notamment positionnée dans l'analyse de données issues d'objets connectés. La startup, basée à Boston, a été créée par un ancien de TechStars Boston, Rohit Gupta. La startup donne l'impression de ne pas avoir grand chose d'autre dans sa besace que ses fondateurs et la capacité à recruter des développeurs sur la côte Est. Elle est très early stage et n'a pas grand chose à raconter à ce stade.

Cette catégorie comprend de nombreux autres acteurs tels que **Alteryx** (2010, $163M), **Predixion Software** (2009, $37M), **RapidMiner** (2007, $36M), **Alpine Data Labs** (2011, $25M) et **Lavastorm** (1999, $55M).

Ecosystème français de l'IA

Depuis 2016, diverses initiatives (France Is AI d'ISAI, le plan France IA du gouvernement) cherchent à valoriser les startups de ce nouveau marché et de mettre en avance l'excellence française et ses opportunités.

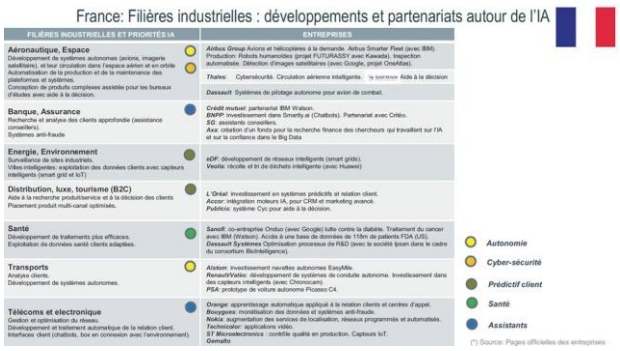
Ce comportement est assez fréquent face à de nombreuses vagues technologiques : dans les jeux vidéos, dans les objets connectés, dans la cybersécurité, pour ne prendre que quelques exemples.

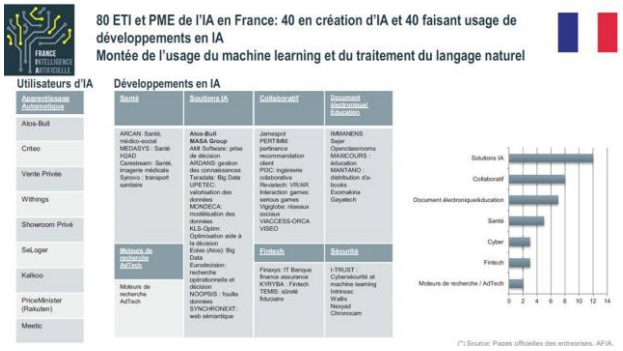
L'habitude est de mettre en valeur voire de monter en épingle l'excellence de nos ingénieurs et de nos chercheurs, qu'ils soient restés en France ou expatriés dans de grandes entreprises du numérique internationales. Et nous avons Yann LeCun, le père des réseaux de neurones convolutionnels, qui œuvre chez Facebook ! ISAI a inventorié de son côté 270 startups françaises opérant dans l'IA dont une bonne vingtaine ont des implantations à l'étranger et notamment aux USA.

Yann LeCun, AT&T, Facebook
Jérôme Pesenti, IBM Watson, BenevolentTech
Emmanuel Mogenet, Google
Thierry Donneau-Golencer, Tempo AI, SalesForce
Antoine Blondeau, Sentient, $140m levés
François Chollet, Keras, Google
David Cournapeau, Scikit Learn, Enthought
Jamal Atif, LAMSADE - Paris
Francis Bach, INRIA, DI-ENS – Paris
Anne-Marie Kermarrec, Mediego
Claude Berrou, IMT Atlantique
Laurence Devillers, CNRS-LIMSI
Isabelle Bloch, LTCI Télécom ParisTech

Raja Chatila, ISIR - Paris
Matthieu Cord, LIP6 - Paris
Jean-Gabriel Ganascia, UMPC-LIP6
Jean-Claude Heudin, Léonard de Vinci
Béatrice Daille, LS2N - Nantes
Sébastien Konieczny, CNRS, CRIL - Lens
Jérôme Lang, CNRS, LAMSADE - Paris
Catherine Pelachaud, CNRS, ISIR - Paris
Henri Prade, CNRS, IRIT - Toulouse
Marie-Christine Rousset, LIG - Grenoble
Marc Schoenauer, INRIA - Paris
Thomas Schiex, INRA - Toulouse
Cordelia Schmid, INRIA - Grenoble
Jean-Philippe Vert, Mines ParisTech - ENS Paris

Il est clair que l'IA constitue une belle opportunité de bien positionner le pays. C'est un secteur prometteur, surtout dans la mesure où les plateformes correspondantes sont encore en devenir. 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 lié à l'IA mais 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.





On a tendance à raisonner en filières industrielles (*ci-dessus*, la cartographie du plan France IA de mars 2017) pour les intégrer, notamment via les Pôles de Compétivité ou dans le cadre de partenariats avec de grandes entreprises françaises que l'on qualifiera de traditionnelles. C'est un énorme piège et une perte de temps. Dans l'IA comme ailleurs dans le numérique, l'intégration est mondiale, pas française. Une startup ne peut gagner à terme qu'en devenant une plateforme. Et les plateformes ne sont pas françaises, 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 et 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.

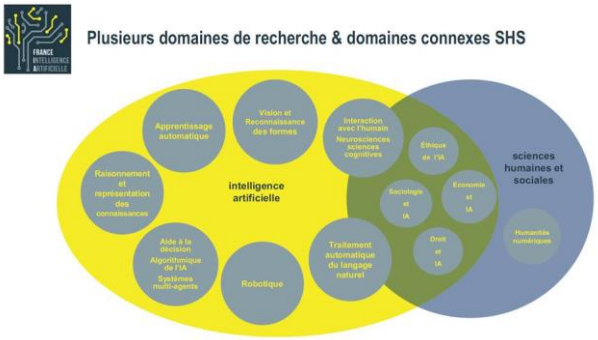
L'autre développement de l'écosystème de l'IA, moins visible, concerne les prestataires de services et ceux qui proposent du service outillé. Nombre d'agences de communication, web agencies et autres entreprises de services numériques se mettent progressivement à l'IA et structurent leurs offres.

Recherche

Avant d'évoquer le cas des startups, faisons 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[218].

La recherche française est disséminée dans plus de 220 équipes de recherche totalisant 5300 chercheurs, avec de nombreux projets collaboratifs associant laboratoires publics, universités et, parfois, entreprise privées. Les principaux organismes se focalisant sur l'IA sont l'INRIA, le CNRS et le CEA.

Les chercheurs français sont 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. Montréal ambitionne ainsi de devenir une capitale de l'IA et a même récemment accueilli un laboratoire d'IA du français Thalès[219].



[218] Cf 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 qui doit être mis à jour par Cédric Villani d'ici début 2018

[219] Cf 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) où travaille la star canadienne de l'IA, Yoshua Bengio, l'IVADO (Institute of Data Valorization), l'Institut d'Intelligence Artificielle du Québec et l'Institut Vector de Toronto.

L'INRIA a publié courant 2016 un excellent livre blanc qui décrit ses priorités et projets dans l'IA[220]. 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.

Le projet Orpailleur mené à Nancy est dédié à la représentation des connaissances et au raisonnement. L'équipe planche sur l'extraction de données dans les bases de connaissances non structurées, et notamment dans le domaine de la santé, le même que celui qui est investi par IBM Watson et plein de startups. Ils collaborent notamment avec le centre de lutte contre le cancer de Nancy.

L'équipe Magnet travaille directement sur le machine learning et l'auto-apprentissage.

Celle de Linkmedia travaille sur l'extraction intelligente de données des contenus multimédias et notamment la découverte de leurs liens et structures.

Le projet Sequel qui est basé à Lille travaille sur l'apprentissage séquentiel de données, notamment celles qui proviennent de l'environnement.

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 vis-à-vis de la société[221].

Une association créée en 1993 fait la promotion de la recherche en IA, l'**Association Française pour l'Intelligence Artificielle** (AFIA). A noter qu'en France, comme il y a toujours du pour et du contre, nous avons aussi une Association contre l'IA, l'AFCIA, créée en 2015 qui vise simplement à interdire à l'échelle mondiale toute recherche sur l'IA[222].

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**[223] 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 réalisa-

[220] Cf Intelligence artificielle, dédis actuels et l'action d'Inria, 2016, 82 pages.

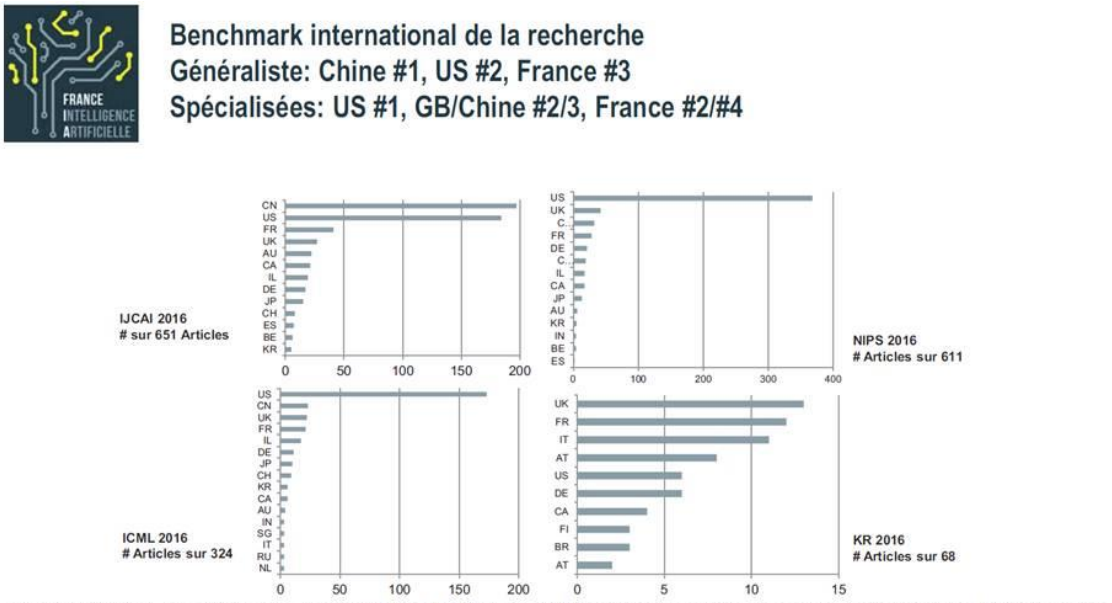
[221] Cela se retrouve notamment dans les travaux de **Laurence Devillers**, du CNRS-LIMSI qui portent sur le langage, sur la détection des émotions et sur l'éthique de l'IA.

[222] Cf On peut être contre l'intelligence artificielle par principe de Irénée Régnauld, publié sur Uzbek&Rica en janvier 2017

[223] Cf le portefeuille de participations d'IT Translation.

tions et leurs applications et que les innovations des startups résultent souvent de la combinaison de plusieurs méthodes et techniques[224].

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.



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[225]. 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.

Startups horizontales

Voici quelques startups que j'ai pu repérer dans les solutions techniques d'IA plus ou moins génériques, en plus de celles qui sont déjà citées dans le corps du document et notamment dans les applications génériques et les applications métiers de l'IA.

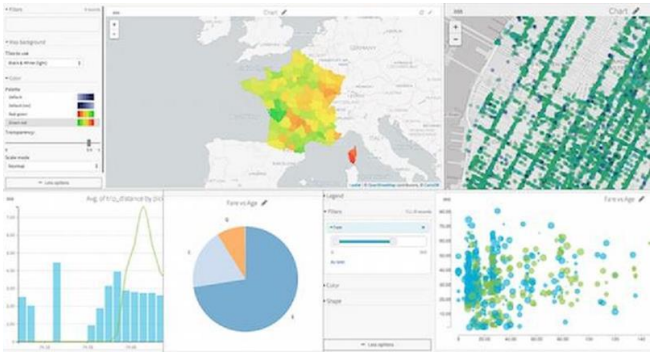
Il subsiste quelques acteurs spécialisés dans la recherche et qui ont intégré petit à petit des techniques d'IA dans leurs offres. Antidot et Sinequa sont anciens dans le pay-

[224] 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.

[225] Cf Meet these incredible women advancing AI research, Topbots, mai 2017.

sage mais, à l'instar de nombreux éditeurs b2b, ils peinent à croître pour atteindre la taille critique, même s'ils commencent à se développer à l'international comme Sinequa qui y réaliserait plus de 50% de son chiffre d'affaire.

Dataiku (2013, $45,7m) fait évoluer les concepts de business intelligence et de data mining avec son Data Science Studio (*ci-dessous*), un ensemble d'outils d'analyse de données qui exploitent du machine learning pour la création de modèles de données et leur simulations. C'est visiblement la startup française la mieux financée dans l'IA à ce jour.



Moodstocks (2008) proposait une solution mobile de reconnaissance d'images, fournie sous la forme d'APIs et d'un SDK multi-plateforme. Elle a été acquise par Google en 2017.

Zelros (2015, 80K€ de love money) propose une plateforme en cloud B2B qui permet aux applications métiers d'accéder aux données structurées ou non ainsi qu'aux modèles prédictifs et en langage naturel via un bot conversationnel exploitable via Slack, par SMS, Skype Entreprise ou équivalents. La startup est basée à Paris.

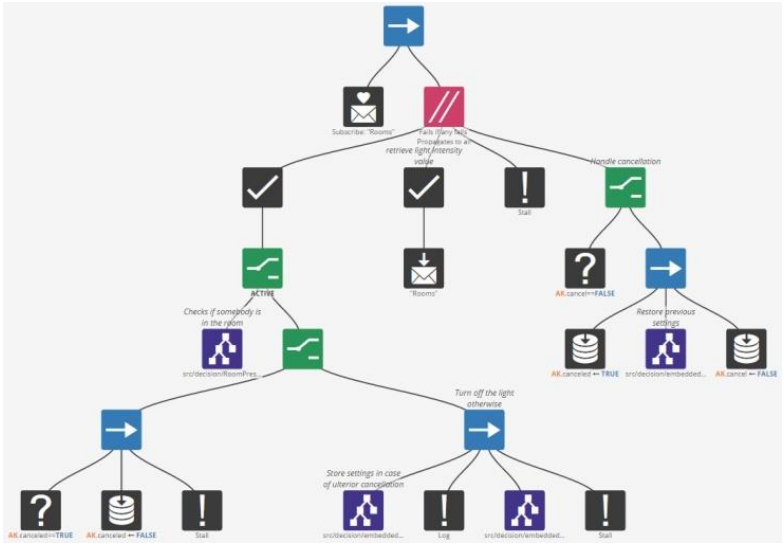
DreamQuark (2014) développe des solutions d'intelligence artificielle à base de réseaux de neurones et de deep-learning avec des mécanismes d'auto-apprentissage capables d'explorer tous seuls tous types de données de les traiter. La startup propose des outils d'analyse via sa plateforme Brain qui permet d'explorer, optimiser et valoriser les données structurées (bases de données) et non-structurées (images, sons, voix) dans les secteurs de l'assurance et la santé.

Objets connectés

C'est un domaine où les entrepreneurs français sont assez prolixes en général. Il n'est donc pas étonnant d'y trouver quelques startups intégrant des briques d'IA dans leurs solutions. Le scénario le plus répandu est lié à la consommation d'énergie et à la maison connectée, avec des solutions faisant de l'auto-apprentissage du comportement de ses habitants pour piloter des actions d'économies d'énergie et d'automatisation diverses.

Craft.ai (2015, $1,1M) 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.



Angus.AI (2014) est un peu l'équivalent de Craft.ai, mais pour les robots et divers objets connectés qui doivent percevoir ce qui se passe dans leur environnement. La startup créée par des anciens ingénieurs d'Aldebaran qui ont développé la partie logicielle des robots Nao et Pepper, propose une solution logicielle embarquée dans les robots leur apportant les fonctions de base de reconnaissance vocale et faciale et de détection d'obstacles. Elles sont fournies sous la forme d'un kit de développement et d'APIs (interfaces de programmation). Ils s'appuient beaucoup sur des solutions open source du marché. Ils travaillent déjà avec la SNCF, mais pas sur des robots.

Ubiant (2011), 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) 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 (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.

Xbrain.io (2012) est une startup française 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. 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 gé-

nération de dialogues réalistes. Son créateur, Gregory Renard, planche sur l'IA depuis près de 20 ans.

Scortex (2015) développe des solutions matérielles et logicielles à base de chipsets neuromorphique en FPGA apportant l'autonomie aux robots et objets connectés et qui intègrent notamment la reconnaissance d'images et de la parole. C'est l'IA de l'informatique embarquée !

Commerce et marketing

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, $3,5M) propose une solution de recommandation prédictive pour les sites de ecommerce qui s'appuie sur de l'intelligence artificielle. C'est un spécialiste du big data marketing. La solution analyse la pondération de la relation entre Internautes et produits et s'appuie sur la théorie des graphes.

Datapred (2014) 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) 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 françaises se positionne sur ce créneau comme **Compellia** (2015), 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, $7,37), 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.

Search'XPR (2013, $2M) est une startup créée à Clermont-Ferrand qui a 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 est mis en œuvre dans la

solution Oorace, destinée au commerce en ligne et même traditionnel. Elle permet 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'appuie 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 est fourni sous la forme d'APIs en cloud. Reste à savoir si les algorithmes relèvent réellement de l'IA et comment ils fonctionnent. C'est la "secret sauce" de la société, vaguement **documentée ici**. Pas forcément de l'IA au sens classique du terme, mais plutôt une algorithmie bien sentie, probablement astucieuse dans sa forme, qui permet d'éviter la force brute de nombreux solutions de machine learning.

Dictanova (2011, $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, $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.

Do You Dream Up (2009) propose une solution de chat automatique pour les sites en ligne. La société est basée à Paris, Bordeaux et Londres. Et elle a déjà une bonne douzaine de clients grands comptes ayant déployé sa solution.

Tastehit (2014) utilise du machine learning et du big data pour personnaliser les sites de e-commerce en temps réel.

CompareAgences (2012) 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.

Cypheme (2015, $1,3M) est une startup proposant une application mobile de détection de produits contrefaits, s'appuyant sur un algorithme de machine learning appliqué à la qualification d'images. C'est une sorte de Shazam de la contrefaçon.

Santé

C'est un domaine très porteur pour les applications de l'IA. Seulement, voilà, nous sommes un peu à la traine dans l'une de ses grandes applications : la génomique. Mais la santé va au-delà de la génomique, heureusement.

CardioLogs (2014) a créé une solution d'interprétation automatique des électrocardiogrammes (ECG) en temps réel s'appuyant sur du machine learning, déjà vue dans la rubrique sur le secteur vertical santé. Uberisation en puissance des cardiologues ?

Pas si vite ! Cela permet surtout de rendre un suivi plus régulier des patients à risques ou atteints de maladies chroniques.

DreamUp Vision (2015) est une startup issue de Dreamquark, une startup spécialisée dans l'analyse de données pour la santé et les assurances. Elle propose une solution d'analyse des images de la rétine obtenues par un fond de l'œil traditionnel. Elle permet de détecter les rétinopathies diabétiques émergentes aussi bien que les ophtalmos. Elle se situe dans un mouvement comprenant quelques autres acteurs dans le monde qui traitent automatiquement les résultats d'imagerie médicale. C'est ainsi le cas d'une autre startup francilienne, **Qynapse** qui analyse de manière itérative les résultats d'IRM cérébrales pour suivre l'évolution de traitements, notamment dans la lutte contre les cancers du cerveau.

Dexstr.io (2014) 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. C'est encore une participation d'IT-Translation.

Applications métiers

C'est là que la créativité est la plus développée.

Riminder (2015) est une startup spécialisée dans les RH qui s'appuie sur du deep learning pour proposer des outils d'aide à la décision. Il aider les chercheurs d'emploi à construire leur parcours professionnel et les actifs à développer leur carrière, en exploitant une base de connaissance de plusieurs millions de parcours de cadres.

White (2015) 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).

niland (2013) est une participation de IT-Translation, la structure de valorisation des projets de recherche issus notamment de l'INRIA. Elle a été créée par des anciens de l'IRCAM et s'appuie sur 10 années de travaux de recherche. Elle utilise le deep learning analysant le contenu de la musique pour rendre son exploration dans les plateformes de diffusion plus intelligente. Elle identifie les similarités entre morceaux pour les classifier automatiquement. La solution sera exploitée par CueSongs (UK, une société fondée par le chanteur Peter Gabriel) et motionelements (Singapour) qui sont dédiés aux professionnels de la musique. La solution est aussi illustrée par le service en ligne **www.scarlett.fm** et s'appuie sur Soundcloud pour vous permettre de créer une web radio personnalisée en fonction de vos goûts.

Dhatim (2008) automatise la gestion des factures et le contrôle des déclarations sociales avec comme premiers clients les opérateurs mobiles (pour les factures) et d'autres (pour les déclarations sociales). Dans ce dernier cas, la solution permet 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 qui déclenche des actions automatisées.

Séline (2013), édité par la société Evi, propose une panoplie d'applications bureautiques intégrant un agent conversationnel permettant de dialoguer et poser des questions en langage naturel. On y trouve notamment un traitement de texte, un tableur, un gestionnaire d'agenda, un carnet d'adresses, un gestionnaire de tâches, une médiathèque, un logiciel de gestion de finances et un gestionnaire de messagerie instantanée. Dilemme classique : faut-il recréer tout un existant complexe pour y intégrer une nouvelle fonction ou ajouter cette fonction aux produits existants du marché (Microsoft Office, Open Office). Question d'ouverture, de simplicité de mise en oeuvre et de modèle économique !

Bayes Impact (2014, $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 connaitre en lançant un partenariat avec Pôle Emploi pour faciliter le rapprochement entre l'offre et la demande d'emplois.

Dans un compte-rendu sur l'écosystème entrepreneurial de La Réunion[226], j'avais aussi identifié quelques startups qui utilisent le machine learning : **logiCells** (ERP sémantique) et **Teeo** (analyse de consommation d'énergie pour les entreprises).

Ce tour est probablement incomplet et les oubliés du secteur se feront immanquablement connaitre. Et d'ici peu de temps, l'usage du machine learning et du deep learning seront aussi courants dans les startups que l'appel à des bases de données NoSQL : une banalité !

Le top du top de la startup d'IA ? Utiliser l'IA dans une solution d'agent conversationnel en cloud qui fait du big data sur des données issues de l'IOT en sécurisant les transactions via des Blockchains. Le Bingo de la startup d'IA est lancé !

[226] Ici : http://www.oezratty.net/wordpress/2016/ecosysteme-entrepreneurial-reunion/.

L'IA dans la 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 bouleversements 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 ?

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 rentrée 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 dans le deep learning, et à une conception à géométrie variable de la notion d'intelligence. 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.

Qui plus est, l'IA conserve un côté magique qui permet de faire prendre des vessies pour des lanternes à de vastes audiences y compris 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. C'est au minimum une grande vague de « logiciels 2.0 » qui est lancée à vive allure.

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.

Les risques

L'IA génère une peur pour l'Homme d'être dépassé par ses propres créations, la peur de perdre le contrôle de son devenir, à la fois intellectuel et pour la maitrise du monde physique.

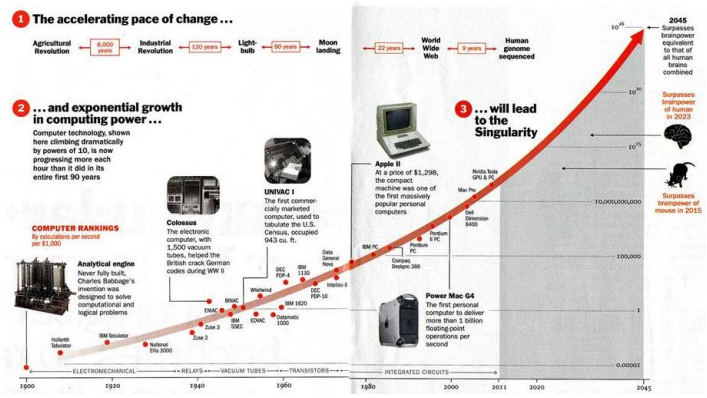
Cette peur est alimentée par la perspective de voir émerger d'ici quelques décennies à peine une IA généraliste quelque peu mythique (AGI = artificial general intelligence) omnisciente, omnipotente et contrôlant tout notre monde physique, et qui pourrait en retour nous asservir.

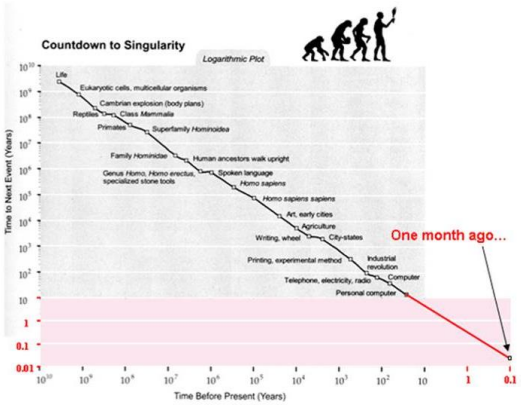
Cette peur s'appuie sur une extrapolation des capacités actuelles de l'IA, une vision un peu simpliste d'une l'application ad-vitam de la loi de Moore et surtout, sur les

méconnaissances des mécanismes de l'IA. On observe ainsi que plus les prévisionnistes de l'IA ont de véritables connaissances scientifiques en IA et en neurosciences, moins ils en ont peur. La plupart considèrent même que l'AGI, tout comme la singularité, sont des mythes[227].

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[228].

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[229].

L'IA présente ainsi 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. 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 dans le futur d'anti-virus spécialisés pour les logiciels de deep learning.

---

[227] 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 !

[228] 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.

[229] Cf 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.

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 systèmes experts) et sur le mimétisme des sens et comportements humains (pour le deep learning). Comme ce qui a été expérimenté involontairement : si un chatbot s'entraine en discutant avec des internautes racistes, il deviendra lui-même raciste. Bref, l'enfer, c'est l'Homme, pas l'IA !

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.

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.

Nous manquons aussi de recul. L'Homme est déjà dépassé par ses propres créations depuis longtemps[230], 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.

Je m'étonne toujours de notre capacité à construire des paquebots et porte-avions de plus de 300 m de long et pesant de 100 000 à 250 000 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 dans la pratique. Une simple clé USB de quelques dizaines de Go peut contenir plus de textes que ce que nous lisons, écrivons, entendons et disons pendant toute notre vie [231]! Et plus les outils numériques stockent l'information et sont faciles à interroger, moins on la retient. Les machines ont toujours été pilotées et contrôlées par l'Homme.

---

[230] 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,

[231] 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 textuelle ne fait probablement qu'à peine quelques Go. Cette évaluation à la louche n'intègre pas la mémoire visuelle et auditive, qui est probablement plus dense.

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[232].

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.

Les pessimistes

La première source de pessimisme 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[233].

Ceci étant dit, les films racontant la vie heureuse de familles 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 le mur de mémorial de la CIA !

[232] 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.

[233] 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 !

Les alertes sur les risques de l'IA gagnent en crédibilité et 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 [234]! Il reprenait à son compte une citation d'Irwin John Good de 1965 publiée dans **Speculations Concerning the First Ultraintelligent Machine**[235] 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 connait bien en trous noirs, il n'est pas forcément spécialisé en réseaux de neurones et deep learning.

9. Conclusions

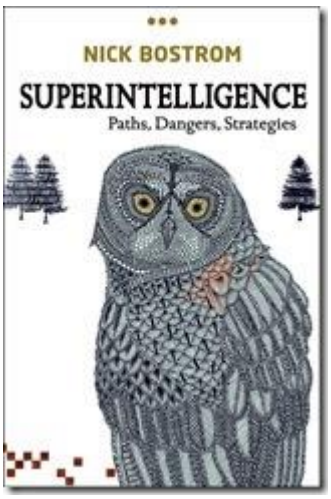
These "conclusions" are primarily the opinions of the writer, as they must be in a paper on ultraintelligent machines written at the present time. *In the writer's opinion then*:

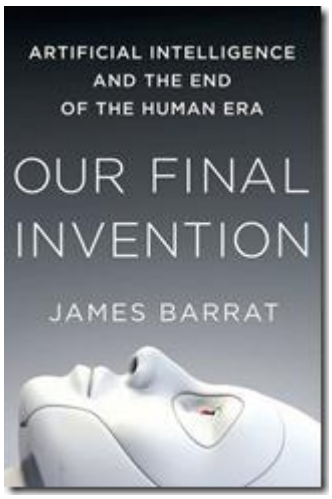
It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make, since it will lead to an "intelligence explosion." This will transform society in an unimaginable way. The first ultraintelligent machine will need to be ultraparallel, and is likely *to be achieved with the help of a very large artificial neural net.*

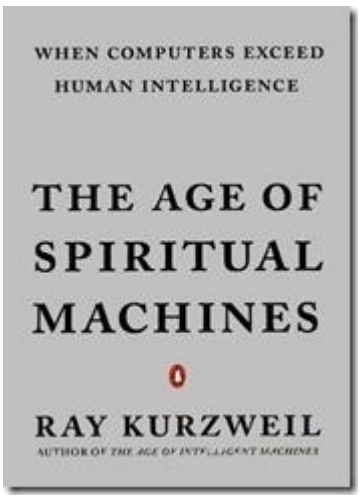
Cette thèse se retrouve décrite dans le menu dans de nombreux ouvrages, comme ceux de Nick Bostrom dans **Superintelligence**, paru en 2014 ou dans **Our Final Invention, Artificial Intelligence and the End of the Human Era** de James Barrat, paru en 2015.

[234] Notamment ici : http://www.bbc.com/news/technology-30290540.

[235] Trouvé ici : http://www.kushima.org/is/wp-content/uploads/2015/07/Good65ultraintelligent.pdf.





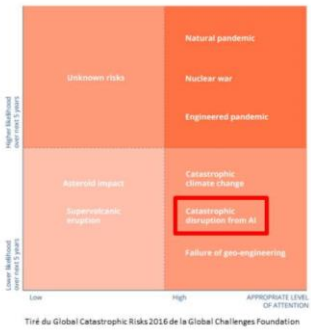


Ces prédictions partent du principe que l'on arrivera un jour à créer une machine superintelligente dont la puissance croitra de manière exponentielle et qui contrôlera toutes nos destinées du fait de l'hyperconnexion des infrastructures physiques et des objets de la vie courante.

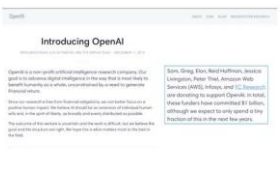
Le cofondateur de Sun Microsystems, **Bill Joy**, avait été l'un des premiers à alerter l'opinion avec **Why the future doesn't need us**, un long texte publié en 2000 dans Wired[236], tirant la sonnette d'alarme sur les dangers des progrès technologiques dans l'IA, les nanotechnologies et les biotechnologies. C'était bien avant la fin du premier séquençage complet du génome humain qui avait coûté une fortune[237].

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**.

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.







[236] Ici donc : http://www.wired.com/2000/04/joy-2/.

[237] On y apprend d'ailleurs qu'il avait rencontré Jacques Attali et que ce dernier avait indirectement influé le cours des événements de Java !

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**, 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 naive du fonctionnement du capitalisme. 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 doit faire de la recherche. C'est un peu comme si une organisation patronale s'était lancée dans une initiative visant à rendre le capitalisme responsable[238]. C'est d'ailleurs la mission que voudraient se donner différents Etats en orientant la recherche et les investissements vers de l'IA responsable dans la lignée de la « tech for good », la technologie pour le bien de l'Humanité et pas celle qui sert à la publicité ciblée qui est sa contraposée la plus courante.

C'est la forme que prend **Partnership on AI**[239], une initiative et association créé en 2016 et rassemblant comme membres fondateurs un bon bout des GAFAMI : Google, Facebook, Amazon, Microsoft et IBM. Il manquait Apple qui les a rejoints en janvier 2017. Donc, ce sont bien exactement les GAFAMI au complet ! L'association est présidée par Eric Horvitz, le patron de Microsoft Research et Mustafa Suleyman de Google, le co-fondateur de DeepMind. Elle doit débattre 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

[238] Cf **OpenAI** dans Wikipedia et **Why you should fear artificial intelligence** paru dans TechCrunch en mars 2016.

[239] Cf https://www.partnershiponai.org/.

réglementaires gênant l'innovation. Eric Horvitz promeut de son côté l'utilisation de l'IA pour le bien de l'humanité[240].

Une autre initiative associe deux fondatrices issus de Microsoft Research et Google : **ArtificialIntelligenceNow**, lancée mi 2016 par 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[241].

Sous des couverts de bonne gestion du principe de précaution, ces initiatives des GAFAMI sont à analyser sous la loupe des pratiques habituelles du lobbying. Elles visent 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 dupes !

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[242], via des nano-électrodes directement implantées dans le cerveau et capables d'activer sélectivement les neurones.

Dans la pratique, 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[243]. 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.

Les 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

---

[240] Cf 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.

[241] Cf 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.

[242] L'idée est inspirée des neural laces de l'auteur de science fiction Iain M. Banks. Cf 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.

[243] 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.

2030-2040[244] en nous promettant monts et merveilles qui peuvent nous encourager à procrastiner sur la résolution des problèmes d'aujourd'hui (réchauffement climatique, surpopulation, inégalités, …).

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.

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[245] et puis **Ginni Rometti** d'IBM qui recommande de ne pas avoir peur des robots[246]. 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[247].

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[248] !

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[249], tourné dans le mauvais sens [250]! 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

---

[244] 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 juste 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. Cf 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 ?

[245] Cf Intelligence artificielle : Zuckerberg (Facebook) n'a pas peur, dans ZDNet, février 2016.

[246] Cf At Davos, IBM CEO Ginni Rometty Downplays Fears of a Robot Takeover de Claire Zillman dans Fortune, janvier 2017.

[247] Cf The optimist's guide to the robot apocalypse de Sarah Kessler, mars 2017.

[248] Mais 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.

[249] 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 ! Cf 'Artificial Intelligence' Has Become Meaningless de Ian Bogost dans The Atlantic, mars 2017.

[250] 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.".

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 »[251]. 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 entrainera à ne pas lui nuire.

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[252].

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[253] soulignant la crédulité de l'écosystème de l'innovation.

Autre méthode, se rassurer une fois encore avec les écrits de **Piero Scaruffi**[254]. 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.

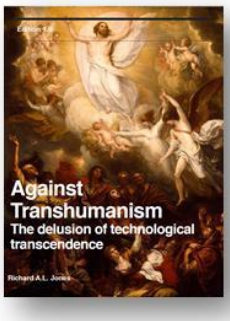
---

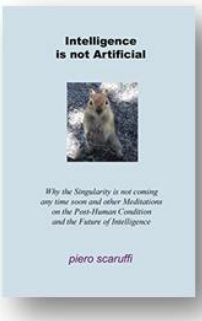
[251] Cf Le mythe de l'IA surhumaine de Rémi Sussan, mai 2017. Kevin Kelly décrit cela lui-même dans Wired en avril 2017 : The myth of a superhuman AI.

[252] Cf 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.

[253] Cf 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.

[254] Comme Demystifying Machine Intelligence.



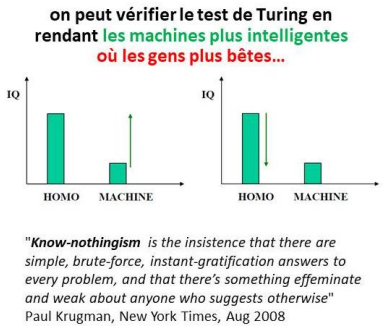


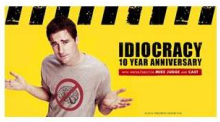
Selon lui, l'Homme a toujours cherché une source d'intelligence supérieure, qu'il s'agisse de Dieux multiples ou unique, de Saints ou d'extra-terrestres. La singularité et les fantasmes autour de l'IA seraient l'incarnation d'une nouvelle forme de croyance voire même de religion, une thèse aussi partagée par Jaron Lanier, un auteur anticonformiste qui publiait **Singularity is a religion just for digital geeks** en 2010[255].

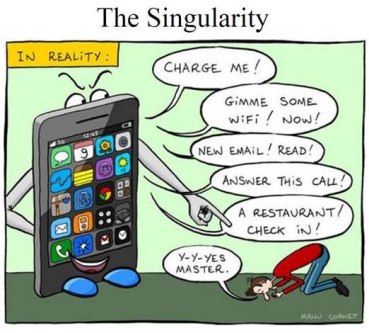
Piero Scaruffi prend aussi la singularité à l'envers en avançant que l'ordinateur pourra fort bien dépasser l'Homme côté intelligence parce que les technologies rendent Homo Sapiens plus bête[256], en le déchargeant de plus en plus de fonctions intellectuelles, la mémoire en premier et le raisonnement en second !

Selon lui, le fait que les médias numériques entrainent 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[257] !

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.







la vie des utilisateurs est déjà largement régie par les logiciels et les plateformes

[255] Ici : http://bigthink.com/devils-advocate/singularity-is-a-religion-just-for-digital-geeks.

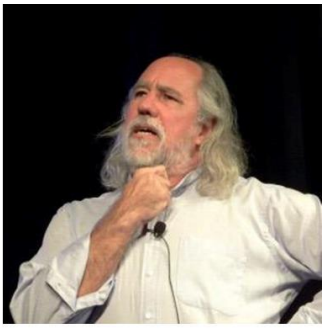
[256] Thèse partagée par **Daniel C. Dennett**, pour qui le véritable danger n'est pas dans les machines plus intelligentes que l'homme mais plutôt dans le laisser-aller de ce dernier qui abandonne son libre arbitre et confie trop de compétences et d'autorité à des machines qui ne lui sont pas supérieures.

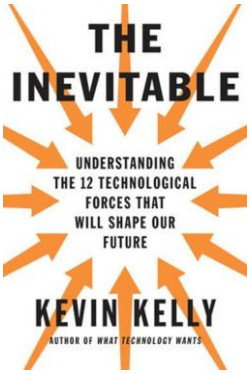
[257] "I am worried that people's attention span is becoming so short that it will soon be impossible to explain the consequences of a short attention span. I don't see an acceleration in what machines can do, but i do see a deceleration in human attention… if not in human intelligence in general", dans Intelligence is not artificial.

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 !

conférence sur les usages "for good" de l'IA


Grady Booch



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.

Transformation des métiers par l'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.

Prévisions de destruction d'emplois

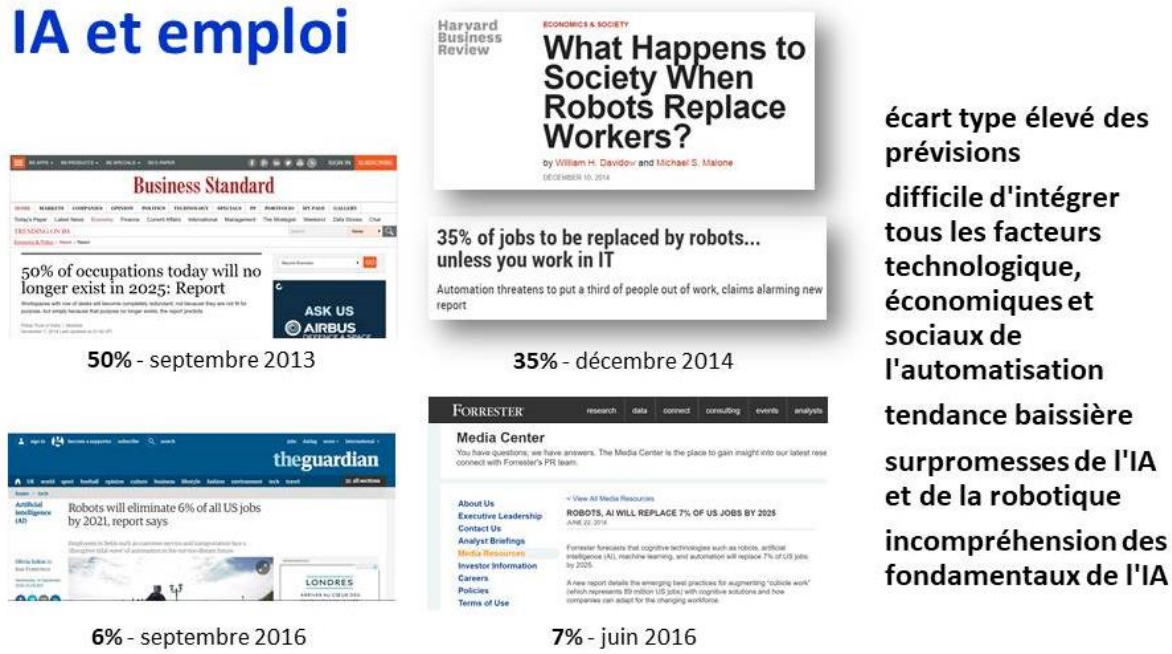
Elles sont plus qu'abondantes ! On y trouve aussi bien de sombres prophéties sur le rôle même de l'Homme dans l'économie que des prévisions plus optimistes, croyant fermement à la destruction-création de valeur schumpétérienne à équilibre positif.

La destruction nette d'emplois liée à l'IA à l'horizon 2023-2015 se situe selon les études entre 6%[258] à 47%[259], avec des prévisions qui suivent une tendance baissière,

[258] 7% pour l'OCDE : OCDE et emplois automatisables, à 7%. Dans The Risk of Automation for Jobs in OECD Countries, mai 2016.

[259] Cf http://www.liberation.fr/evenements-libe/2016/05/09/dans-20-ans-plus-de-60-des-metiers-sont-amenes-a-disparaitre_1451563.

la principale prévision de 47% datant de 2013[260] et celles de 6% à 7% datant de 2016 Ca donne une belle marge d'erreur et de manœuvre !

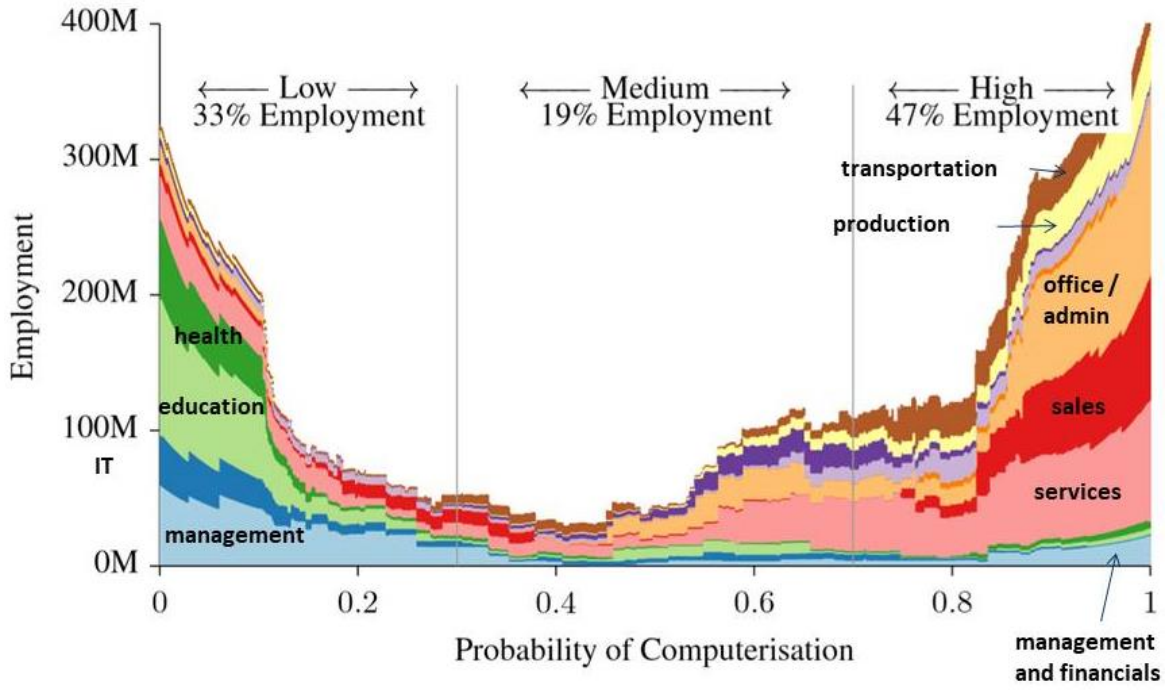


Cela illustre que les tendances lourdes sur le marché de l'emploi, si elles auront bien lieu, interviendront un peu plus tard. Pour que tel ou tel emploi disparaisse d'ici 5 ans, il faudrait que les technologies correspondantes soient disponibles aujourd'hui compte-tenu de l'inertie du marché, des parties prenantes, des budgets et des déploiements. Si elles ne sont pas encore disponibles, il faudra alors attendre plus de 5 ans pour qu'elles aient un impact sur l'emploi ! 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.

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.

[260] L'étude « à 50% » est anglaise : The Future of Employment: How susceptible are jobs to computerisation? Publiée en 2013, elle segmente avec plus de précisions que la moyenne 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 rique d'automatisation (>60%) et surtout dans les services, la vente et l'administratif.

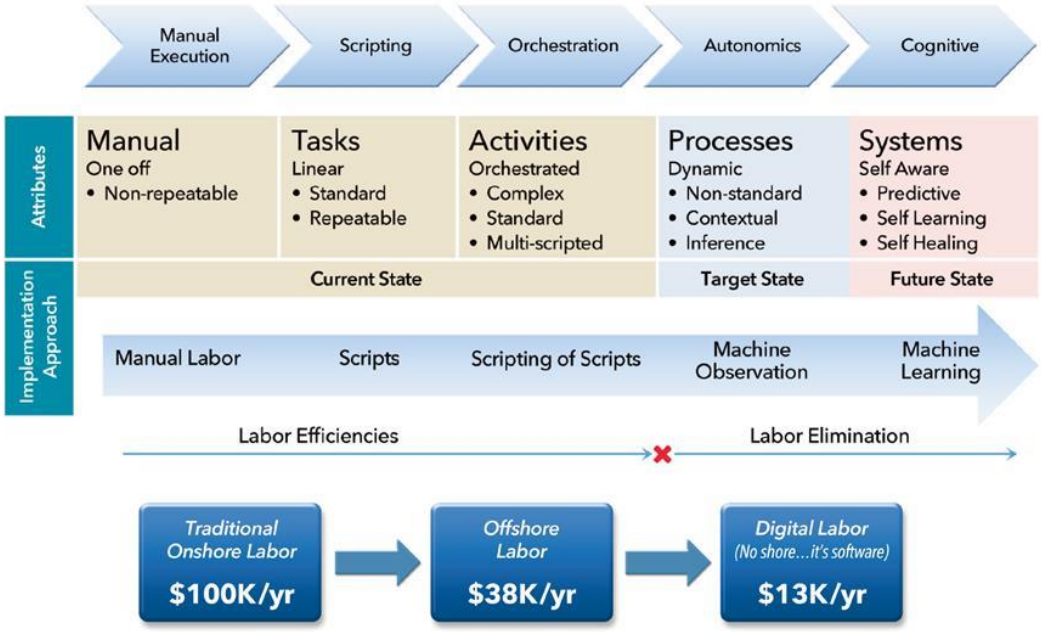


Aujourd'hui, sont en ligne de mire prioritaire de transformation ou remplacement par les technologies numériques et par l'IA :

- Les **métiers du passé** : les ouvriers de lignes d'assemblage (déjà remplacés par des robots), les caissiers (remplacés en partie par des automates de self-service) et les centres d'appels de taxis (remplacés par des applications mobiles et de la commande vocale).
- Les **métiers en train d'être automatisés** : surtout les conducteurs professionnels, qui seront remplacés à moyen terme par des véhicules à conduite autonome, suivis de nombreux métiers de services, notamment dans les professions libérales administratives et dans la finance qui est de plus en plus automatisée[261]. 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[262].

[261] Cf 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.

[262] Cf Introduction to Robotic Process Automation, a primer, de l'IRPA, Institute for Robotic Process Automation



- Les **métiers qui seront automatisés plus tard**, totalement ou partiellement comme dans la santé, le management, l'audit et même dans la recherche[263]. Certains prospectivistes vont jusqu'à prévoir que le métier de développeur va lui-même être automatisé, ce à quoi je ne souscris pas[264].

La principale erreur d'appréciation de ces prévisions est qu'elles confondent 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.

[263]Cf 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.

[264] 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.

Il faut aussi 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 émotionnels[265], 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 guerrre mondiale, ils sont dominants dans ces catégories (tourisme, transports, médias, publicité, services divers, boutiques de tatouages, etc).

Il faut aussi adopter une vue globale des métiers. Certes, certains métiers seront de plus en plus automatisés ou rendus plus efficaces via l'automatisation. 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[266]. Des phénomènes de vases communicants peuvent intervenir. Telle disparition entraîne la création d'emploi dans des secteurs connexes voire entièrement différents des métiers disparus.

---

[265] Apporter de l'amour en plus de l'IA ! Cf A blueprint for coexistence with artificial intelligence de Kai-Fu Lee, juillet 2017.

[266] Exemple : http://www.engadget.com/2016/05/04/autonomous-robot-surgery/.

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 de l'émotionnel et des loisirs. On voit cela dans le rapport **Artificial Intelligence and Robotics and Their Impact on the Workplace**[267] qui évoque comme métiers du futur, les data scientists[268], les métiers de créatifs, les freelances et les métiers manuels dont les services à la personne. C'est aussi l'approche de l'économiste français **Nicolas Bouzou**[269], adepte de la prise de recul historique sur les craintes de destruction de l'emploi pour nous rassurer.

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.

Ensuite, on se trompe souvent sur le terme et même la nature des chamboulements. Surestimés à court terme, sous-estimés à long terme, mais surtout mal appréhendés dans leur réalité technique et économique.

Ainsi, dans **Les robots veulent déjà nous piquer notre job**[270] 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[271] 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.

---

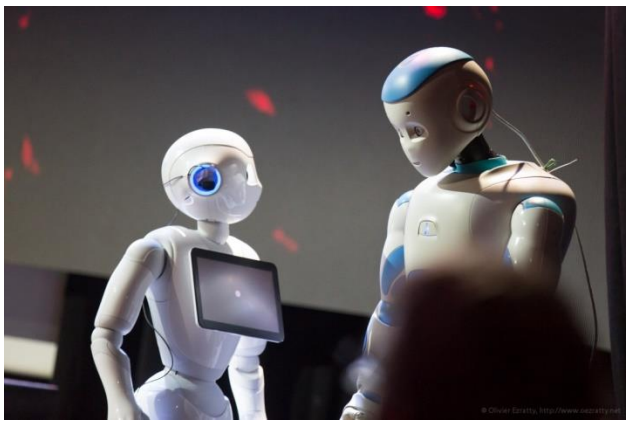
[267] Artificial Intelligence and Robotics and Their Impact on the Workplace produit par l'Iinternational Bar Association Global Employment Institute en avril 2017, 120 pages.

[268] Ce qui constitue une vision très réductrice des métiers techniques nécessaires pour faire tourner de l'IA !

[269] Noramment dans « Le travail est l'avenir de l'Homme », 2017.

[270] Cf http://www.presse-citron.net/les-robots-veulent-deja-vous-piquer-votre-job/.

[271] Mes photos du robot Pepper dans la boutique Softbank de Tokyo en 2014 sont ici : http://www.oezratty.net/wordpress/photos/?ws=geECJj.



Les robots Pepper et Romeo d'Aldebaran Robotics (groupe Softbank) en apparente discussion, pendant l'événement des 10 ans de Cap Digital à Paris fin mars 2016. En fait, ils ne discutent pas vraiment. Et un shutdown avait bloqué le premier pendant de longues minutes. Work still in progress !

En y regardant de près, l'étude en question est un article publié dans The Conversation, **Are robots taking our jobs**[272]. 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**.

Autre exemple légèrement exagéré : celui du fonds d'investissement **Deep Knowledge Venture** de Hong Kong et Londres qui a nommé en 2014 un logiciel d'intelligence artificielle à son board, dénommé VITAL[273] ! Il devait servir à identifier les projets les plus prometteurs dans la santé, l'une des spécialités de ce fonds couvrant aussi le big data, les fintechs et l'IA en général. Evidemment, le relai de cette annonce a donné lieu à quelques exagérations : le logiciel est ainsi facilement passé de membre du board à CEO de l'entreprise[274]. On n'est plus à une exagération près pour forcer le trait ! Mais c'est comme si on disait que Excel est à la tête des entreprises, ce qui n'est d'ailleurs pas si faux que cela dans pas mal de cas d'un point de vue symbolique ! Au passage, on ne peut que nommer des personnes physiques dans ces rôles-là, même à Hong Kong [275]! Les effets d'annonces de ce genre sont rarement suivis de preuves et retours d'expérience ! On peut par contre observer la récursivité du modèle. Le fonds a en effet investi dans **Transplanetary**, une startup spécialisée dans la recommandation d'investissements dans les industries spatiales, avec son logiciel SPOCK (Space Program Ontologically Computed Knowledge)[276].

En tout cas, 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

---

[272] Cf https://theconversation.com/are-robots-taking-our-jobs-56537.

[273] Cf http://www.itbusiness.ca/blog/hong-kong-vc-firm-appoints-ai-to-board-of-directors/48815.

[274] Ici donc : https://humanoides.fr/2014/05/une-intelligence-artificielle-a-la-tete-dune-entreprise/.

[275] Cf https://en.wikipedia.org/wiki/Deep_Knowledge_Ventures.

[276] Mais le site de Transplanetary ne fonctionne pas et la startup n'a pas de fiche dans la Crunchbase.

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 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. 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.

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 50000 robots avaient été déployés[277], 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 prédictions sont rarement analysés. S'y mêlent allègrement la science-fiction, la science et la fiction.

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 [278]!

Ces livres 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

[277] Cf http://www.generation-nt.com/foxconn-foxbot-robot-assemblage-humain-actualite-1914702.html.

[278] Cf http://www.oezratty.net/wordpress/2015/derive-exponentielles-3/.

(sélection directe à l'arrivée) des centres d'appels, les chatbots (qui peuvent nous rendre aussi rapidement fous que les SDA), 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[279]. 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).

Du côté de la vision macro-économique, la majorité de ces ouvrages ont 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 des fonctionnaires qui sont souvent les derniers a être robotisés car protégés par la lenteur de l'innovation dans les administrations et le manque de courage politique.

Ces livres font aussi peu de cas de prédictions sur le devenir du système financier. Ils indiquent qu'il est à l'origine de la concentration de la richesse sur les plus fortunés, qu'il détourne la valeur ajoutée des salaires vers le capital, et qu'il pousse à l'automatisation, faisant courir l'économie à sa perte. Et pourtant, le système financier est basé sur un point clé, bien mis en avant par Yuval Harari dans l'excellent **Sapiens** qui relate de manière très synthétique les dynamiques de l'histoire de l'humanité : le système financier, surtout celui des prêts, repose sur la confiance dans le futur. Cette confiance est la clé de voûte du capitalisme et du système financier. Or cette confiance est en train de s'effondrer pour tout un tas de raisons (emploi, dette, environnement, …). Cela se retrouve dans les faibles taux d'intérêts actuellement pratiqués. Si la robotisation met à genoux le système financier et l'économie derrière, des freins naturels se mettront peut-être en place. Ou pas, car le pire est toujours probable !

Autre manque de prédiction : l'impact des progrès issus de l'IA sur la démographie ! 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 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[280].

D'ailleurs, la majorité des études sur l'impact de l'IA sur le futur de l'emploi ne s'intéressent qu'aux pays développés. L'impact de l'IA sur les payés émergents est rarement abordé et il 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

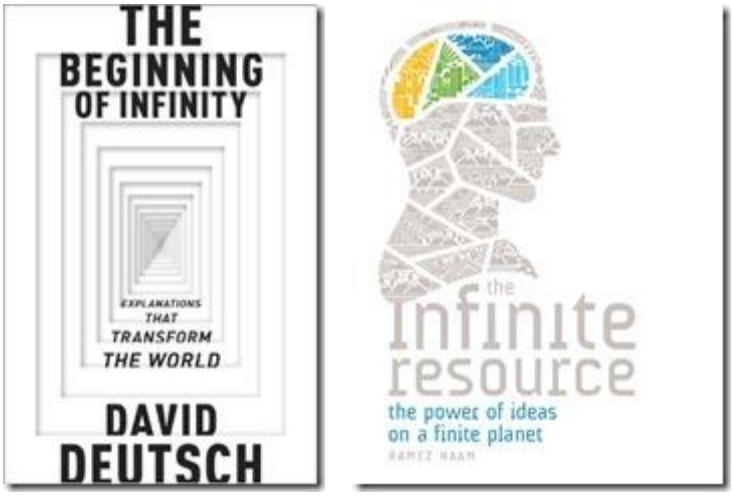
[279] Le service n'est pas vraiment plus rapide mais on économise le temps de transport vers un point de service ou de vente.

[280] 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.

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.

La *Ludditisation* des métiers n'est généralement pas évoquée par les prévisionnistes, 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 !

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.



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.

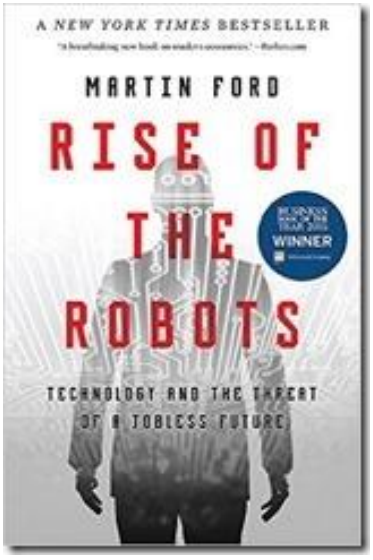
Autres lectures et points de vue

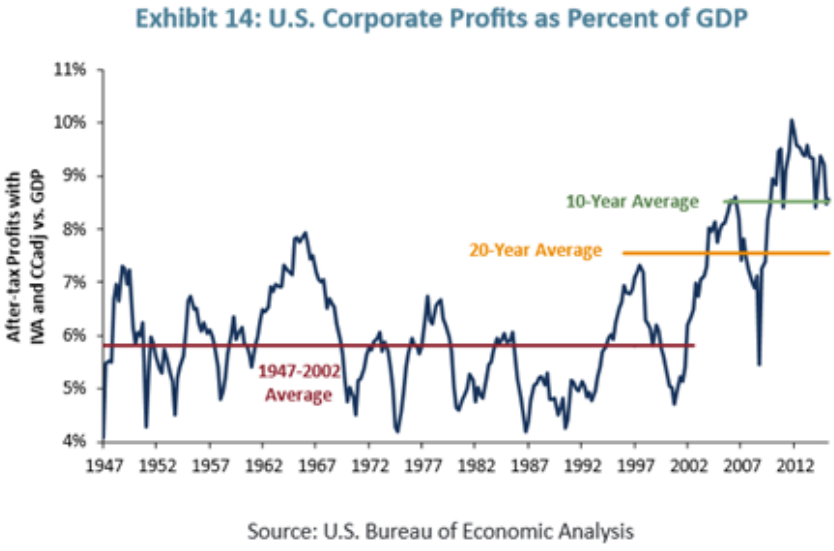
Cette revue éditoriale sur la robotisation des métiers et le futur des emplois n'est pas terminée.

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[281] 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 finan-

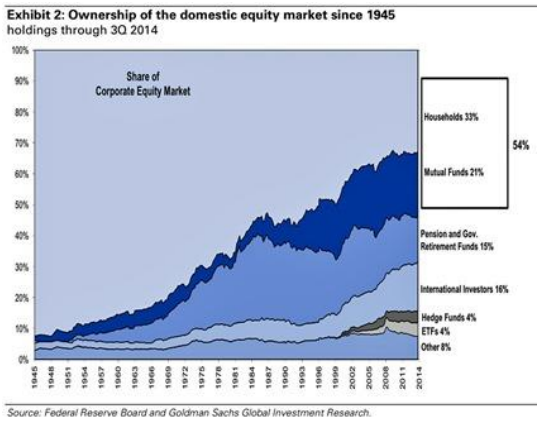
[281] Cf http://www.educationanddemocracy.org/FSCfiles/C_CC2a_TripleRevolution.htm.

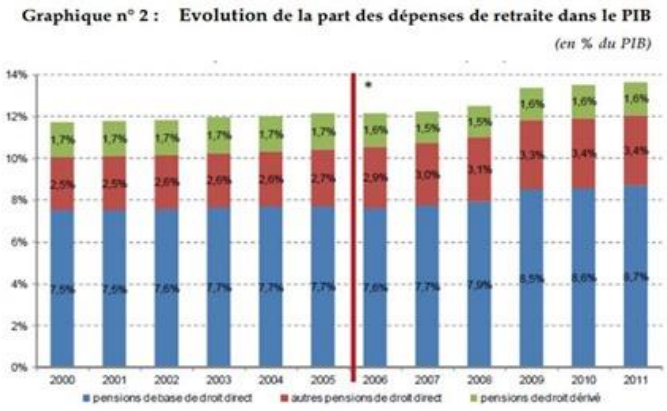
cement des nouveaux investissements technologiques. Les technologies sont devenues un facteur d'inégalité au profit des technologues et des détenteurs de capital, tout du moins aux USA. La "finance" réalloue aussi les profits au bénéfice des plus riches. Plus un pays a un système financier développé, 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, 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, aussi pour alimenter des fonds de retraite. Les méchants fonds spéculatifs (hedge funds) ne représentent que 4% de l'actionnariat des entreprises américaines [282]!

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

[282] Voici la **source** du schéma correspondant.

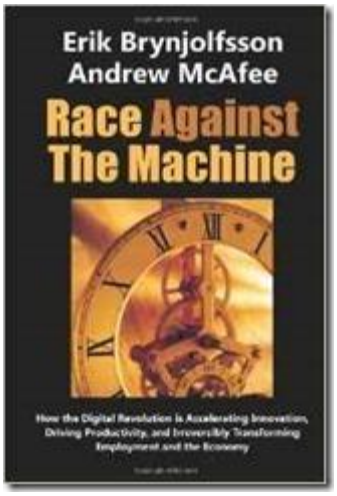
par le congrès US créerait une inertie souhaitable pour protéger les emplois non qualifiés. En même temps, elle favorise l'offshore de métiers qualifiés en plus des métiers faiblement qualifiés qui sont déjà externalisés à l'étranger. Et la politique immigatoire encore plus restrictive de Donald Trump 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.

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. 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 émergent 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 a été ni voté, ni encore appliqué à fortiori !

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.

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.

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 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.

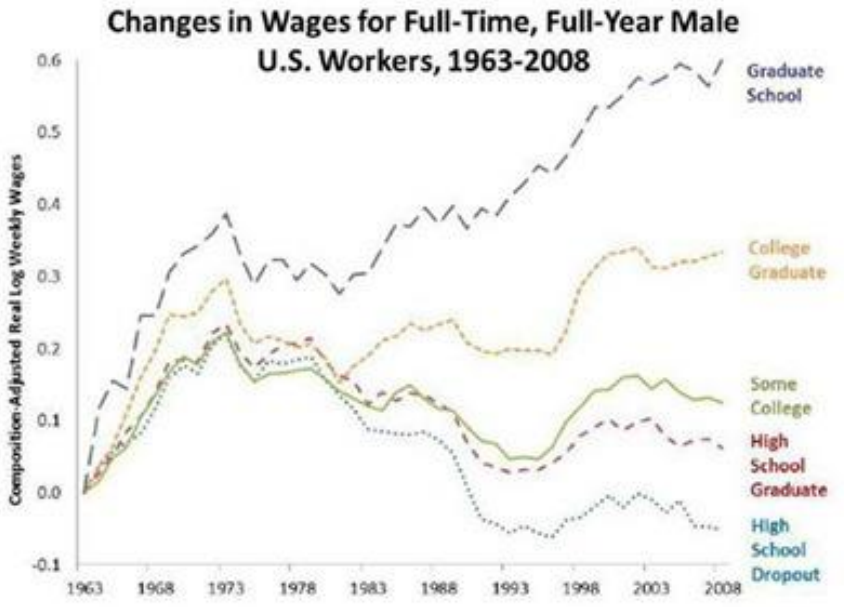


Figure 3.5: Wages have increased for those with the most education, while falling for those with the least. Source: Acemoglu and Autor analysis of the Current Population Survey for 1963-2008.

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. Donc, associer la créativité et la maitrise 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 recommandent 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.

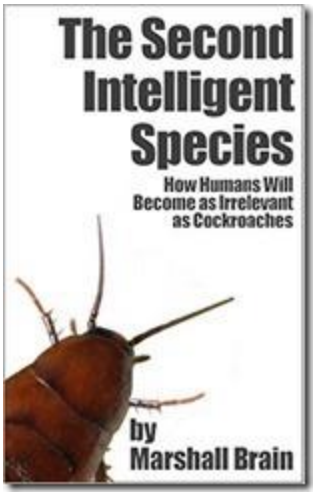
- Côté **lois et fiscalité**, ne pas alourdir la législation du travail. Rendre les embauches plus attractives que la robotisation des métiers au niveau des charges sociales et taxes, ce qui rappelle une bonne partie de la politique de l'emploi en France, qui ne nous réussit pas si bien. Ne pas réguler 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.

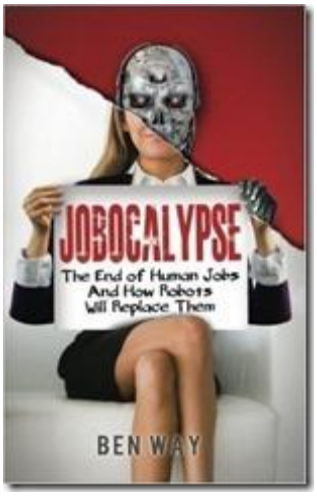
C'en est presque un plan "à la Macron" : favorisons l'entrepreneuriat et tous les problèmes sociétaux se régleront d'eux-mêmes. Un peu trop classique !

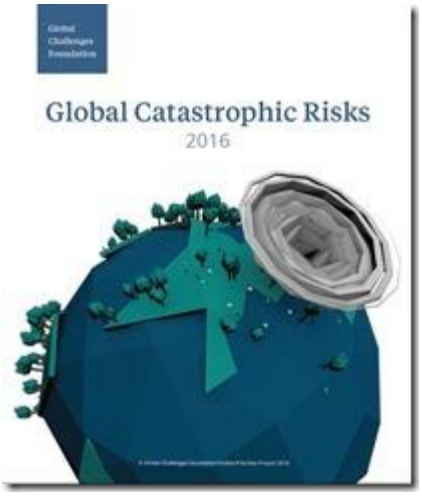
The Second Intelligent Species: How Humans Will Become as Irrelevant as Cockroaches (2015), de Marshall Brain, grossit le trait en annonçant que les scientifiques sont en train de créer une seconde espèce intelligente, les robots et l'IA, qui va nous dépasser et supprimer la majorité des emplois. Les 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. Faut voir…

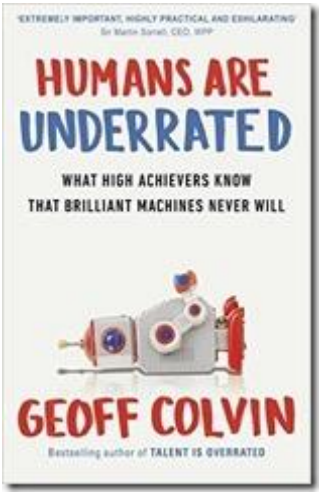






Jobocalypse (2013) de Ben Way, que je n'ai pas lu (désolé…), part du principe que nous sommes *déjà* envahis par les robots et que la disparition d'emplois liée à l'automatisation est une histoire ancienne. Il anticipe que même les métiers les plus qualifiés seront remplacés par des robots car ils s'autoalimenteront. 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[283], 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.



L'étude **AI, Robotics and the future of jobs** du Pew Research Center, parue en 2014[284], recense de son côté l'avis de divers spécialistes dont certains estiment que la

[283] Ici : http://www.globalchallenges.org/reports/Global-Catastrophic-Risk-Annual-Report-2016.pdf.

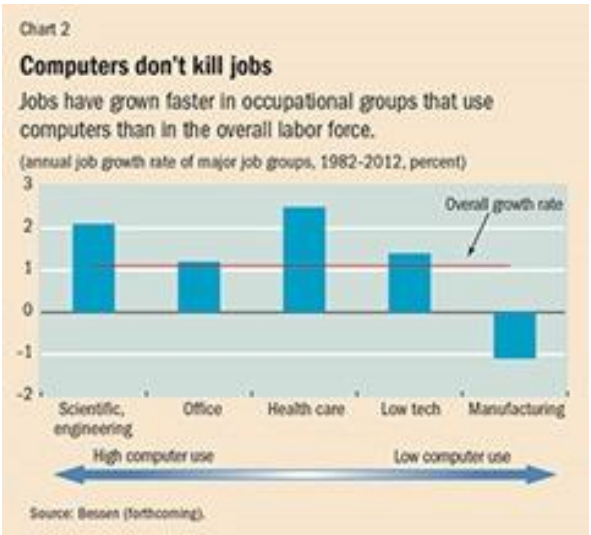
[284] Ici : http://www.pewinternet.org/files/2014/08/Future-of-AI-Robotics-and-Jobs.pdf.

moitié des emplois sont menacés à l'horizon 2025. 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.

C'est aussi l'avis de **Darrell M. West** de la Brookings Institution dans **What happens if robots take the jobs**[285] qui prévoit des créations de jobs dans plein de secteurs et des disparitions dans peu de secteurs. On retrouve cette thèse dans **Toil and Technology – Innovative technology is displacing workers to new jobs rather than replacing them entirely** (2015) de James Bessen[286], d'où sont extraits les schémas ci-dessous qui montreraient que les ordinateurs ne sont pas à l'origine de la suppression d'emplois.



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

---

[285] Ici : http://www.brookings.edu/~/media/research/files/papers/2015/10/26-robots-emerging-technologies-public-policy-west/robotwork.pdf.

[286] Ici : http://www.imf.org/external/pubs/ft/fandd/2015/03/bessen.htm.

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[287].

Elle était reprise par Dov Seidman dans **Harvard Business Review** en 2014[288] (*ci-dessous*). C'est une belle conclusion, même si frisant quelque peu l'utopie[289].


Harvard Business Review

ECONOMY

From the Knowledge Economy to the Human Economy

by Dov Seidman

NOVEMBER 12, 2014

In the human economy, the most valuable workers will be hired hearts. The know-how and analytic skills that made them indispensable in the knowledge economy no longer give them an advantage over increasingly intelligent machines. But they will still bring to their work essential traits that can't be and won't be programmed into software, like creativity, passion, character, and collaborative spirit—their humanity, in other words. The ability to leverage these strengths will be the source of one organization's superiority over another.

Comment éviter de se faire robotiser

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 !

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.

---

[287] 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 ». !

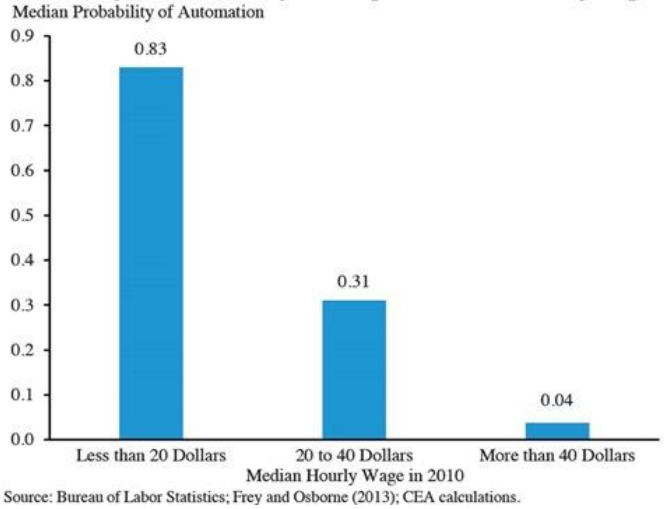
[288] Ici : https://hbr.org/2014/11/from-the-knowledge-economy-to-the-human-economy.

[289] 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.

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, maitriser à la fois des métiers traditionnels et les technologies numériques qui peuvent les transformer. Malheureusement, les sciences et technologies n'attirent pas tant que cela les jeunes et notamment en France, comme une enquête mondiale récemment réalisée par **Randstad**[290] le montre.

Figure 5-15
Probability of Automation by an Occupation's Median Hourly Wage



Source: Bureau of Labor Statistics; Frey and Osborne (2013); CEA calculations.

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**[291] 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[292], 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. 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

[290] Voir http://www.influencia.net/fr/actualites/tendance,etudes,francais-sur-quatre-conscient-etre-remplace-par-robot,6296.html.

[291] Cf https://www.whitehouse.gov/sites/default/files/docs/ERP_2016_Book_Complete%20JA.pdf.

[292] Voir http://www3.weforum.org/docs/WEF_FOJ_Executive_Summary_Jobs.pdf.

dont le contenu émotionnel sera le plus dense, comme expliqué dans **Les 10 compétences clés du monde de demain**[293].

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[294].

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 et Watson n'utilisent 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 que Watson 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 !

On apprend souvent du passé pour (mieux ?) décider du futur. Mais de nouveaux éléments complexifient la donne. Par exemple, doit-on faire une analogie entre la montée des populismes et démocratures dans le monde et la situation des années 1930 ou d'avant la première guerre mondiale ? Qu'est-ce qui est similaire et qu'est-ce qui est différent ? Comment anticiper la dimension émotionnelle qui secoue un peuple ? Quand est-ce que le peuple est au bord d'une révolte ? Comment l'anticiper ? 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 ?

---

[293] Ici : https://missphilomene.com/2016/03/20/les-10-competences-cles-du-monde-de-demain/.

[294] Ici : http://www.fastcompany.com/3058800/the-future-of-work/how-to-avoid-being-replaced-by-a-robot.

Je m'étais aussi demandé en 2013[295], 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 !

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[296]. 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[297].

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**

---

[295] Cf http://www.oezratty.net/wordpress/2013/technologies-et-assassinat-jfk/.

[296] Cf http://watson2016.com/.

[297] 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 ici : http://www.oezratty.net/wordpress/2016/origine-consequences-trump/.

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 !

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[298], et qui dépasse les générateurs de pipeau déjà bien connus. 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[299] ! 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 !

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.

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 appor-

[298] Cf Political Speech Generation de Valentin Kassarnig.

[299] Ici : https://github.com/valentin012/conspeech.

terait un peu de rationalité et serait capable de prendre des décisions non basées sur le côté obscur des émotions[300].

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[301] 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 de la politique fiction. Ca ne saurait tarder vue l'imagination débridée des scénaristes ! Un « House of Card » avec un « Special Assistant to the President » qui soit une IA à commande vocale ferait son effet !

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

Dans la pratique, l'IA est en fait déjà utilisée en politique dans quelques contextes sporadiques :

- Le ciblage de prospection terrain réalisé lors de la campagne de Barack Obama en 2012.
- Pour la segmentation et le ciblage d'audience clés pour une élection comme ce que l'anglais **Cambridge Analytica**[302] a permis de faire en analysant 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

[300] Cf Should Politicians be Replaced by Artificial Intelligence? Interview with Mark Waser, 2015.

[301] POTUS = President Of The United States.

[302] 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.

pour y arriver ! Il suffit d'exploiter les technologies existantes et d'avoir de mauvaises intentions. Le mal, c'est l'Homme !

- L'exploitation sémantique des analyses terrains réalisées par les volontaires de En Marche, par la startup française **Proxem**. Proxem 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[303]. 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.

Tout cela reste encore artisanal. Mais ce n'est peut-être que la face visible d'un gros iceberg méconnu.

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 à diverses menaces encore hypothétiques.

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.

Mais l'IA est aussi une révolution industrielle et les Etats ont compris qu'il ne fallait pas louper le coche, sans forcément disposer de recettes miracles.

Des deux côtés de l'Atlantique, les Etats sont des plus prolixes en rapports et plans autour de l'IA.

Barack Obama a été interviewé par Joi Ito dans Wired en août 2016 et articulait déjà une vision claire des enjeux autour de l'IA[304]. 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

[303] Cf http://mashable.com/2017/04/04/predictgov-artificial-intelligence-congress.

[304] Cf https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/. Sa position est bien documentée et articulée. On se met à fantasmer sur une interview comprenant exactement les mêmes questions, mais posées à l'actuel POTUS, Donald Trump. L'A/B testing de POTUS serait saisissant !

pas remplacer les hommes, et l'usage des véhicules autonomes devrait être régulé. S'en suivit Artificial Intelligence, Automation, and the Economy en décembre 2016, qui anticipe 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.

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 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).

Le plan France IA

Je faisais un tour d'horizon de ce plan et rapport de 350 pages dans un article publié en mars 2017[305] et dont voici un résumé légèrement actualisé.

Le diagnostic d'ensemble de ce plan faisait état d'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'intelligence artificielle. Il évoquait le rôle stratégique des données qui alimentent l'IA.

On y trouve surtout un état des lieux très fouillé, largement exploité dans ce document, qui comprend une cartographie de la recherche française en intelligence artificielle, complété par celle des entreprises l'utilisant. Il comprend 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 détaille aussi les stratégies de grands acteurs américains et chinois ainsi que les plans lancés par différents gouvernements dans le monde. C'est un inventaire très utile pour se faire une idée de la concurrence mondiale.

Le plan se focalise sur sept secteurs d'activité : la construction automobile, la relation client, la finance, la santé, les énergies renouvelables, la robotique et l'éducation nu-

[305] Que j'ai commenté ici au moment de sa publication : http://www.oezratty.net/wordpress/2017/les-hauts-et-les-bas-du-plan-france-intelligence-artificielle/.

mérique. Il comprend des dizaines de recommandations génériques, ou dans certains seulement de ces secteurs d'activité comme pour les véhicules autonomes.

Le plan d'action prévoyait notamment :

- La préparation d'une 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 qui n'est pas sans rappeler l'équivalent proposé dans le cadre du plan France Génomique 2025 annoncé en juin 2016.
- La création d'un consortium public-privé sur l'intelligence artificielle.

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.

Ce plan avait été réalisé en seulement deux mois, ce qui est un record de rapidité. Il n'était qu'une première étape d'un processus qui sera probablement itératif. Il avait le mérite d'avoir rassemblé autour de plusieurs tables un bel échantillon des acteurs français impliqués de près ou de loin dans l'IA. Il se poursuivra notamment dans le cadre de l'initiative France IA.

Une mise à jour de ce plan a été confiée à Cédric Villani de l'Assemblée Nationale, par Mounir Mahjoubi, l'actuel Secrétaire d'Etat en charge du numérique. Cédric Villani est l'actuel président de l'Office Parlementaire d'Evaluation des Choix Scientifiques et Technologiques qui est une entité commune entre le Sénat et l'Assemblée Nationale.

Le 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[306] « **Pour une intelligence artificielle maîtrisée, utile et démystifiée** » complétait le plan France IA de l'exécutif en abordant surtout la dimension sociétale et réglementaire.

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

[306] 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.

CNRS (GREYC, IRIT, LAAS, LAMSADE, LATTICE, LIF, LIG, LIMSI, LIPN, LIRMM, LORIA, LRI, TIMC) totalisant environ 300 chercheurs permanents. C'est à peu près l'effectif de DeepMind, la filiale de Google, mais dont les équipes ne font pas que de la recherche. Sur les nombreuses questions d'éthique et juridiques, le rapport fait le résumé d'un grand nombre d'autres rapports, notamment étrangers.

Le rapport est en tout cas agréable à lire et fort bien documenté avec de nombreux renvois sur des écrits de référence, qu'ils viennent de rapports divers ou d'ouvrages de référence dans le domaine de l'IA.

Il comprend surtout une quinzaine de propositions regroupées en trois parties reprenant les éléments du titre du rapport :

Côté **maitrise**, 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 maitrise 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[307] et aussi de promotion de la diversité et de la place des femmes dans la recherche en IA[308] (il faudrait ajouter, dans l'ensemble de son écosystème). Bref, des propositions tout à fait censées. Le rapport préconise, sans précision, la création de projets de recherche transversaux structurants, en confondant peut-être ce qui ressort de la recherche fondamentale qui est par nature éparpillée de la recherche appliquée et de la création de solutions technologiques qui est naturellement plus transdisciplinaire. De tels projets doivent répondre à des défis en mode « top bottom » comme ceux de la DARPA sur la conduite autonome en 2004. Le rapport préconise la création de champions européens sans élaborer de scénario : top-bottom à la Airbus/Ariane ou bottom-up par le développement du tissu des startups.

---

[307] 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.

[308] 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* ».

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 (qui pourrait déjà exister avec VivaTechnologie).

Dans le détail et sans surprise, la rapporteuse Dominique Gillot (PS) préconisait la création d'une cotisation sociale assise sur les agents autonomes et les robots, sans préciser le salaire de référence, et l'autre rapporteur, Claude De Ganay (LR) y était opposé.

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 demi-douzaine d'années, cela donnerait du grain à moudre. Supporter une augmentation moyenne de 50% du chômage n'est pas facile à absorber, même si des pays comme la Grèce et l'Espagne sont déjà passés par là pour d'autres raisons.

Quelles politiques économiques adopter ? Elles tournent toujours autour du développement économique avec les moyens économiques à disposition des états 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**[309], le rapport annuel 2016 sur l'économie de la Maison Blanche publié en janvier 2017 à la fin de l'administration Obama, on découvre qu'aux USA et en 2013, les startups ont créé 2 millions d'emploi 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 est dédiée à la défense ! Et au milieu des années Reagan, elle en représentait les deux tiers ! Cela explique pourquoi tant de projets autour de l'IA sont financés par la DARPA. 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[310], 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 va-

---

[309] Ici : http://www.presidency.ucsb.edu/economic_reports/2017.pdf.

[310] Cf le transcript exact de son intervention : https://www.cnbc.com/2017/05/23/read-the-full-transcript-of-cnbcs-interview-with-treasury-secretary-steve-mnuchin.html.

lu 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.

Joi Ito, Scott Dadich, and President Barack Obama photographed in the Roosevelt Room of the White House on August 24, 2016.

Op-Ed Trump's Treasury secretary is an Artificial Intelligence denier

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[311]. 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.

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[312]. 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 l'agriculture

311 Cf https://www.technologyreview.com/s/602747/todays-artificial-intelligence-does-not-justify-basic-income/ et https://medium.com/france/les-secrets-bien-gard%C3%A9s-du-revenu-universel-9f8e2cb6a841#.jwc20flg9.

312 Ici : http://fortune.com/2017/02/25/bill-gates-robot-tax-automation-jobs/. Voir diverses réactions : https://www.bloomberg.com/view/articles/2017-02-28/what-s-wrong-with-bill-gates-robot-tax, http://www.businessinsider.com/bill-gates-robot-tax-brighter-future-2017-3?IR=T.

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. 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. Sans compter le fait qu'aujourd'hui, diverses études montrent que les pays les plus robotisés sont ceux qui se développent le mieux !

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.

Il est plus simple de bien taxer les profits des entreprises. 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.

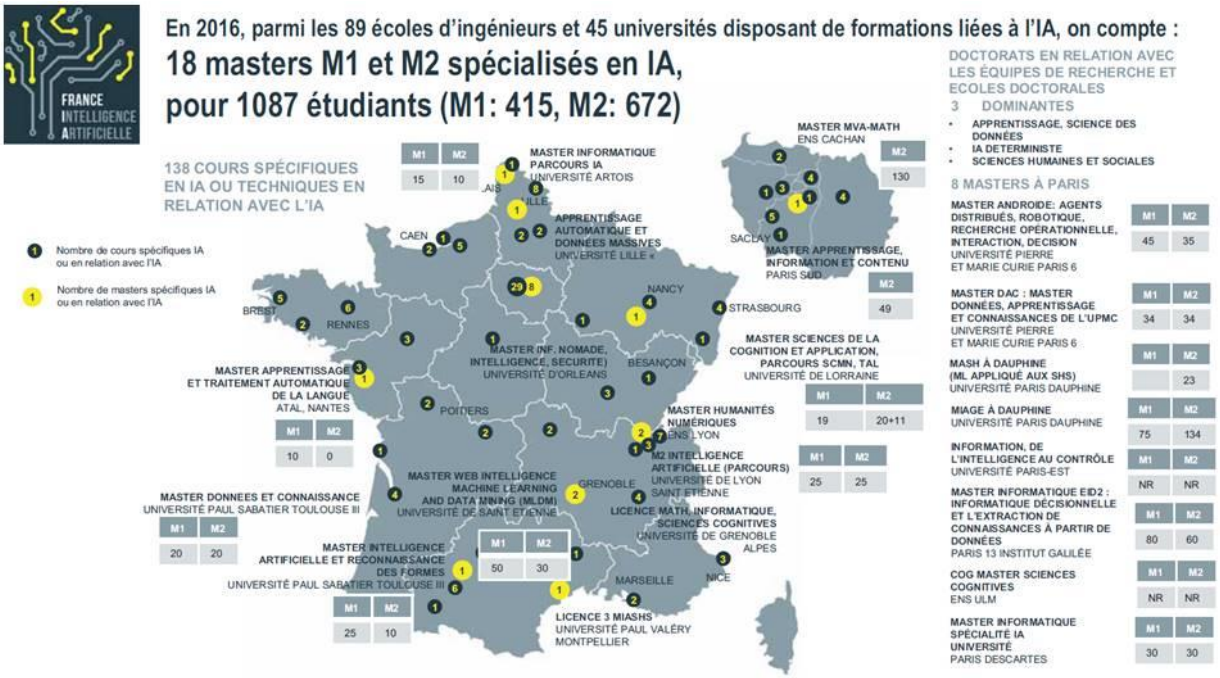
Qui plus est, le risque de pertes d'emplois lié à l'automatisation est plus fort dans les métiers non physiques que dans les métiers physiques. Un expert-comptable est plus menacé par l'IA qu'un kinésithérapeute ou une sage-femme ne le sont par des robots. Dans le cas de la robotisation dématérialisée, 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 linguis-

tique 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.

Education

Le Rapport France IA fait un inventaire des formations dans l'IA. Les Masters Spécialisés ne produisent 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. Et nous avons vu qu'IDC prévoit que dans deux ans, plus de la moité des développeurs devront maitriser l'IA.



Le décalage est évidemment énorme et irréaliste, mais il montre que la situation va être rapidement tendue. 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 claquent des doigts.

D'où les propositions 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).

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 récent « 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'ingénieur 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. Mais pas du jour au lendemain.

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.

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 en France). 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 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.

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 entraine 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[313]. 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.

La **personnalité juridique des robots** a déjà été évoquée au niveau de l'Union Européenne. L'avocat Alain Bensoussan (*ci-contre*) suggère, avec enthousiasme[314], de créer un véritable droit des robots, situé entre le droit des biens et des personnes[315].

S'y ajouterait la création de référentiels robotiques aux niveaux éthiques, culturels et normatifs.

[313] Il me semble que c'était dans cette intervention : « Inside the NSA: An Evening with General Michael Hayden », datant de 2014.

[314] Cf 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.

[315] Alain Bensoussan a aussi lancé divers servcies juridiques en ligne basés sur de l'IA avec sa propre équipe de développeurs. Cf https://www.alain-bensoussan.com/avocat-intelligence-artificielle/. Il propose notamment une solution de justice prédictive.

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[316] 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.

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 GDPR, la nouvelle réglementation européenne, d'ici mai 2018, les entreprises européennes vont devoir 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.

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 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 réentrainer tout le réseau à partir de zéro pour éviter que celui-ci reconnaisse un utilisateur en fonction de son comportement ?

Se pose aussi une question cornélienne de stratégie économique. Les entreprises européennes ne peuvent pas facilement collecter les données personnelles alors que leurs concurrentes américaines et chinoises n'ont pas les mêmes limitations. L'économie européenne est donc handicaptée dans la quête de la dominance dans les applications de l'IA.

S'y ajoute la **règlementation de l'usage des véhicules autonomes**, surtout pendant leur longue phase de cohabitation avec les véhicules traditionnels.Va-t-on faire passer leur permis à des IA et comment ? Et se pose la question de la fameuse gestion des dilemmes lorsqu'un véhicule autonome doit choisir entre deux formes d'accidents et de dommages corporels.

Quid sinon de l'application du Premier Amendement qui régit la liberté d'expression aux USA, à des robots logiciels ?

---

[316] 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.

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 !

L'IA dans l'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, comment, avec qui et pour obtenir quels résultats.

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.

Discours

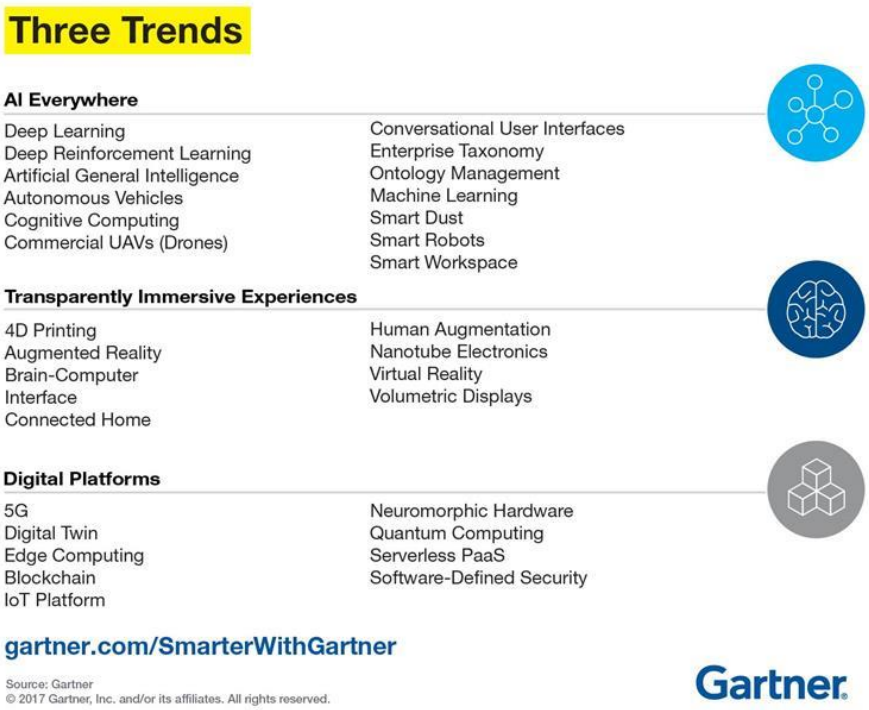
En à peine deux ans, l'IA est devenue la tendance numéro 1 du numérique, alimentée par les performances médiatiques de Google, DeepMind et IBM Watson. L'effet de suivisme est patent chez tous les cabinets de conseils et d'analystes qui ont tous leurs livres blancs sur l'IA, générique ou sectorielle.

Nombre de ces livres blancs sont lénifiants, rappelant les définitions de l'IA (machine learning, deep learning, vision, langage, …) et présentant quelques vagues études de cas marginales. Cet exemplaire ci-dessous issu d'**Infosys** en Inde est édifiant, notamment du côté d'une présentation qui aurait bien pu tenir sur deux pages au lieu de 14[317], et qui ne dit donc pas grand-chose de bien original.

Selon le **Gartner**, l'IA était la première des trois grosses tendances en 2017[318], 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 (qui sont loin d'être adoptables par les entreprises en l'état actuel).

[317] Cf More power to the energy and utilities, from AI, Infosys, 2017.

[318] Cf Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017, août 2017.



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 actuel[319]. En 2015, **General Electric** prévoyait que les objets connectés génèreraient $15T de croissance en 20 ans. **Cisco** évaluait cette croissance à $14 ,4T[320] en moins de 10 ans. On flaire un peu de double booking dans ces prévisions mirobolantes !

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.

Les bénéfices potentiels sur le plan financier sont évidents. L'Intelligence Artificielle ne permettra pas seulement de gagner du temps, elle permettra aussi de décupler les

[319] Cf AI to drive GDP gains of $15.7 trillion with productivity, personalisation improvements, juin 2017.

[320] Cf The Internet of Everything is the New Economy, septembre 2015.

potentialités des entreprises. En améliorant la qualité de vie et la satisfaction au travail des employés, l'IA permettra à ces derniers d'être plus productif et de se focaliser sur l'innovation.

Pour prospérer, les entreprises doivent se préparer à l'Intelligence Artificielle dès maintenant. Le personnel doit être formé en conséquence, des investissements doivent être réalisés et une collaboration rapprochée avec des experts doit être initiée. Les entreprises devront impérativement comprendre comment, où et pourquoi faire appel à l'IA. L'Intelligence Artificielle n'a rien d'une technologie clé-en-main. Elle requiert des spécialistes compétents, à la fois en Intelligence Artificielle et dans le champ d'activité propre à l'entreprise.

Vous remplacez l'IA par les objets connectés ou le cloud et le discours serait quasiment le même !

Vous avez aussi **Avanade** dont le rapport **Technology Vision 2017**[321] 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.

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**[322] (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'entrainement, des mathématiques, des patterns et beaucoup d'itération avec de l'intervention humaine.
- **L'AI est réservée à une élite technologique** et pour les GAFA. Dans la pratique, l'IA est exploitable par toutes les entreprises, notamment via les nombreuses ressources disponibles dans le cloud.
- **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.

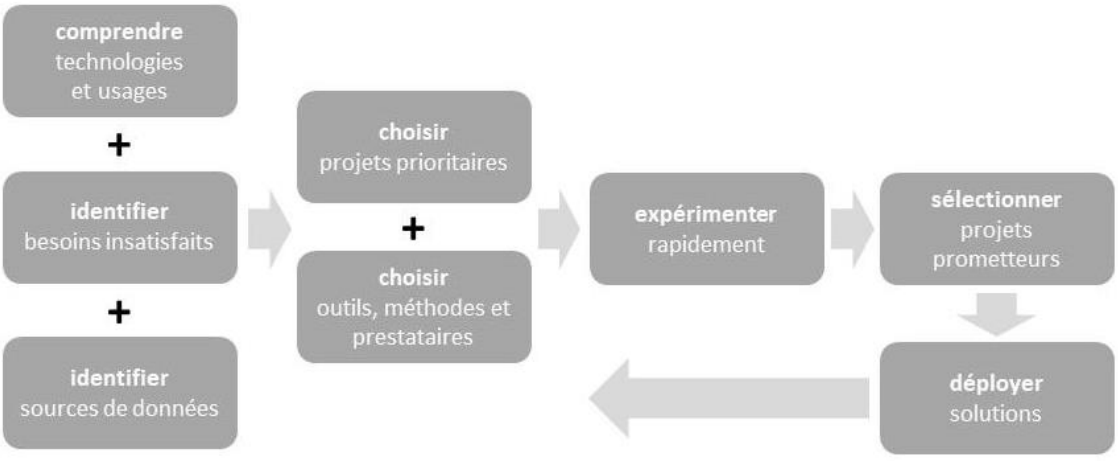
[321] Cf Get ready for the AI first world, 2017.

[322] **CrowdFlower** (2007, $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.

- **Les algorithmes sont plus importants que les données**. L'expérience montre que le contraire est plutôt vrai. Ceci étant, la performance des algorithmes joue un rôle clé dans la qualité des résultats dans le deep learning, et surtout dans leur performance, notamment la rapidité de la phase d'entrainement des modèles.
- **Les machines sont meilleures que les Hommes**. Non, car les machines ont presque toujours besoin d'interventions humaines. Leur intelligence est alimentée par l'expérience et l'intelligence humaines. Qui plus est, l'expérience montre que des IA couplées à des hommes sont supérieures aux IA seules. Et enfin, les hommes et les machines n'ont pas les mêmes capacités et se complètent.
- **Les machines vont remplacer les Hommes**. Dans la pratique, les machines augmentent les capacités humaines et réciproquement.
- **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 des moteurs de règles. 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

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 clients, des inefficiences connues de l'organisation, et des potentialités technologiques. Il faut donc avoir comme bagage de départ une certaine compréhension de ce que les différentes briques et outils de l'IA pourraient apporter. Il faut aussi en connaitre les contraintes actuelles, techniques et économiques.



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 alors reboucler la boucle pour améliorer les projets déployés et découvrir la potentialité de nouvelles technologies d'IA apparues depuis le début du cycle.

Les cycles de développement vont aussi évoluer. Dans l'IA, la mise au point des modèles d'entrainement de machine learning et de deep learning génère des allers et retours plus long que la correction de bugs classiques. L'entrainement 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.

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[323].

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.

[323] 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.

- **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.
- **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.
- **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.
- **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.
- **Développer une communauté autour de l'IA et échanger** : idem, comme sur tous les sujets technologiues du moment (IOT, VR, BlockChain), l'IA étant d'ailleurs souvent un outil associé à ces différents domaines.
- **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'entrainement et dans les paramètres des réseaux de neurones entrainés. Il est important de bien conserver les jeux de données d'entrainement, 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.

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.

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.

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. 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 entrainer 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.

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. Et la majorité des solutions logicielles sont open source.

Les acteurs se rémunèrent avec du service, des solutions métiers payantes, des ressources en cloud, voir du matériel spécialisé.

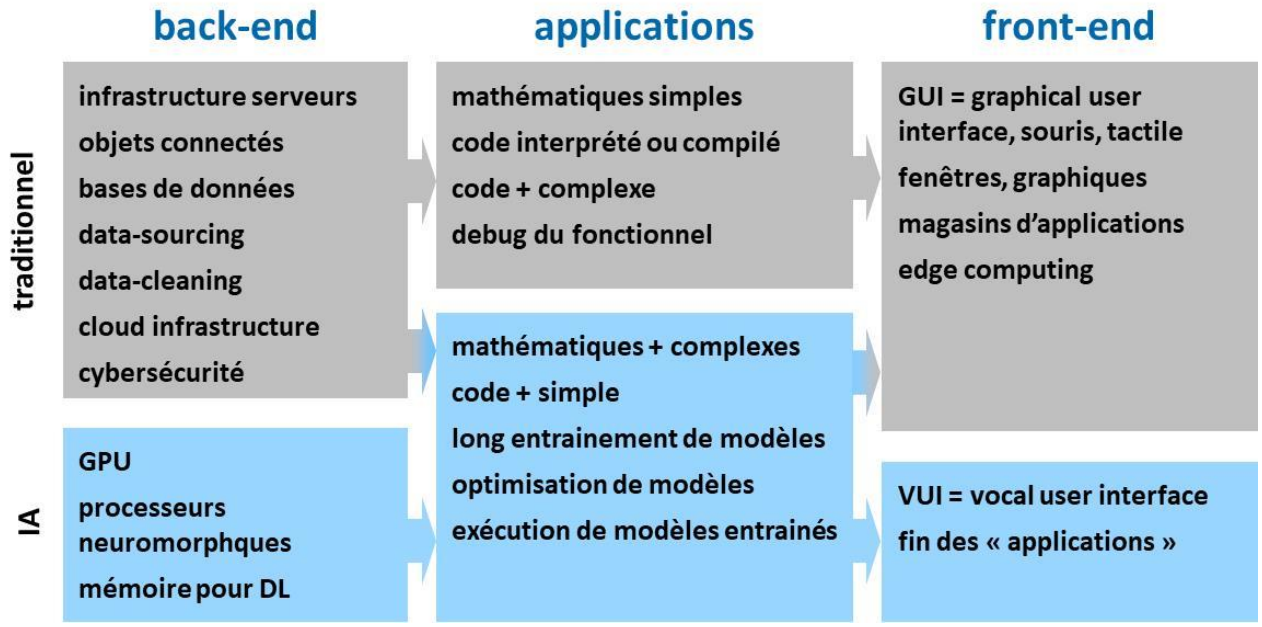
Il va falloir déterminer où exécuter ses solutions. Si elles sont demandeuses de ressources machine importantes, avec des serveurs à base de GPU, il sera censé de faire appel à des fournisseurs de telles ressources en cloud. Et des fournisseurs qui sont évidemment à même de protéger les données de l'entreprise, même les jeux de tests. Ils existent et 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.

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. La qualité des données alimentant les outils de l'IA. En aval, les applications propo-

sées 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.



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.
- **Spécialistes métier** : que l'on retrouve habituellement dans les fonctions de MOA (Maitrise 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.

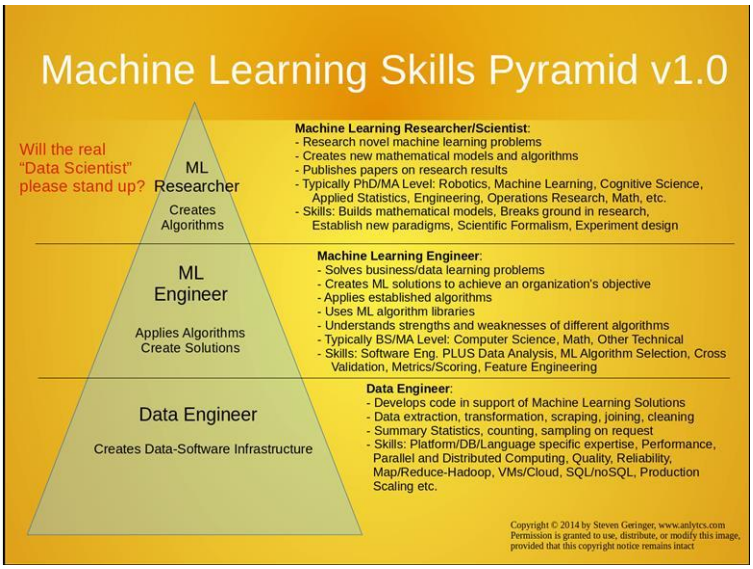
- **Concepteurs d'interactions avec les IA et robots** : qui maitrisent l'utilisation de données comportementale et l'ergonomie pour concevoir et spécialiser les interfaces avec les utilisateurs moins qualifiés et les clients. Ils ont des compétences en design !
- **Entraineurs d'IA** : moins ou pas 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.

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.

des métiers qui se structurent en permanence

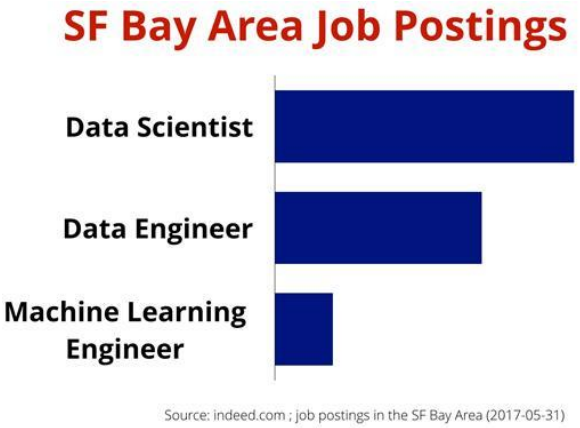


Après avoir rencontré les pires des difficultés à recruter de bons développeurs, les entreprises de services, les éditeurs de logiciels, les entreprises utilisatrices tout comme les startups vont rencontrer de grandes difficultés à identifier les bons talents à même de paramétrer un moteur de deep learning[324] !

[324] L'Américaine **Kaggle** (2010, $16m), acquise 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.

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.

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. Dans les mois et années à venir, les grands acteurs du service vont probablement faire l'acquisition de petits acteurs spécialisés dans l'IA.

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).

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[325] ! Les équipes existantes peuvent et doivent s'emparer de l'IA :

[325] Point de vue partagé dans Qui doit gérer la stratégie en entreprise, de Robin Ferrière, Orange Business Services, septembre 2017.

- Les **DSI** pour intégrer leur dimention technique, l'urbanisation du système d'information, le lien entre les applications de l'IA et le legacy, les données et les infrastructures.
- Les équipes de **maitrise 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.

Epilogue

Nous voici au terme de ce petit voyage 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.

Il est en effet difficile de faire le tri entre la mode, les effets de manche, les annonces tonitruantes et les applications pratiques de l'IA. Des généralisations et des extrapolations abusives sont souvent construites autour de performances médiatisées comme la victoire d'AlphaGo au jeu de Go ou celle de Watson à Jeopardy. Le phénomène enfle, maintenant que l'IA est devenue un argument marketing générique pour les startups comme pour les grands groupes du numérique. Et les prédictions vont bon train, de la fin des métiers à la fin de l'Homme lui-même.

Après ce tour, vous aurez saisi que les prouesses récentes de l'IA sont liée à des progrès parallèles : dans les méthodes et algorithmes qui s'améliorent continument, dans le matériel et dans les données qui les alimentent. 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 !

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 les objets connectés. Il faut évidemment se garder de faire des généralités, comme lorsque l'on présente la BlockChain comme la poudre de perlimpinpin universelle de l'économie de la confiance.

Nous avons pu découvrir de nombreuses startups françaises dans le domaine de l'IA, aussi bien au niveau des techniques horizontales que des applications métiers. Nous avons aussi des talents français également établis aux USA ou dans des entreprises américaines, comme Yann LeCun qui a créé le laboratoire d'intelligence artificielle de Facebook à Paris. Reste à transformer cela en avantage stratégique et en emplois ! Bâtir une Startup Nation de l'IA, voilà un beau défi qui s'annonce ! 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.

Une opportunité existe pour bien positionner la French Tech sur ce créneau porteur qui structurera vraisemblablement les plateformes numériques des années à venir. Comme d'habitude, il s'agit d'être les premiers à créer des plateformes mondiales de grande ampleur, pas juste de créer des myriades d'applications métiers disparates. L'excellence en R&D ne se traduit pas nécessairement en innovations et réussites économiques sinon, la France serait championne du monde des industries numériques depuis des décennies !

Enfin, 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. 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. Les deux exemples les plus connus sont l'empire Ottoman qui a mis trois siècles à adopter l'imprimerie à caractères mobiles après son invention par Guttenberg ou la Chine qui a brutalement bloqué ses échanges maritimes aux débuts du 15ième siècle.

Qui plus est, les prévisions des prévisionnistes n'engagent que ceux qui y croient. Elles sont souvent à côté de la plaque. 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 créer des solutions qui, certes, répondent à des besoins 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 capitalistique 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.

A.I. Magazine : un site d'information en anglais sur l'IA. http://ai-magazine.com.

AI Magazine : un magazine US sur la recherche en IA édité par l'Association for the Advancement of Artificial Intelligence. http://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. http://aiplaybook.a16z.com/

Chatbot Magazine : un site US sur les chatbot. https://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. http://www.jair.org/.

Journal of Machine Learning Research, http://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. http://maisouvaleweb.fr/.

Nanalyze est un bon site web faisant le tour de l'actualité autour des startups de l'IA. http://www.nanalyze.com.

Singularity Hub : magazine de l'actualité scientifique teinté par les technologies dites exponentielle, https://singularityhub.com/

Voicebot : un autre magazine US sur les chatbots. https://www.voicebot.ai/.

Dictionnaire anglais/français de l'IA

Anglais	Français	Commentaire
Back propagation	Rétro-propagation	Technique d'apprentissage de réseau de neurones
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.
Feature map		Utilisée dans les convnets
GDPR : General Data Protection Regulation	RGPD : Règlement Général de Protection des Données	Nouvelle régulation européenne de protection des données personnelles applicable à partir de mai 2018.
Filters	Filtres	Utilisée dans les convnets
Machine learning	Apprentissage machine	
Neuromorphic chipsets	Composants neuromorphiques	Processeurs spécialisés pour les réseaux de neurones.
Neuronal 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	Pourrait avoir des applications futures dans le deep learning.
Recurrent neuronal networks	Réseaux de neurones récurrents	Type de réseau de neurones pour le deep learning.
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.
Uncanny valley	Vallée de l'étrange	Phénomène se manifestant lorsque l'on est mal à l'aise face à un robot humanoide trop réaliste.

Glossaire

AGI : Artificial General Intelligence, IA de niveau équivalent à celle de l'homme. Tout du moins dans la capacité de raisonnement.

Alexa : service en ligne d'agent conversationnel d'Amazon, fonctionnant par reconnaissance vocale et intégré dans son objet connecté Echo.

Algorithmes génétiques : algorithmes s'améliorant d'eux-mêmes par un processus d'évolution voisin de celui du vivant, avec techniques de croisements.

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 supérieur à celle de l'homme.

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'entrainement 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'entrainement. 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.

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 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 entrainement 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: 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.

Google Now : agent conversationnel de Google, fonctionnant sous la forme d'une application mobile.

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.

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'entrainement 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.

Watson : nom de l'ordinateur d'IBM ayant gagné au jeu Jeopardy en 2011 et mettant en jeu des agents conversationnels évolués, appliqués dans différents métiers comme dans la cancérologie. Watson est depuis devenu une plateforme logicielle avec un ensemble d'interfaces de programmation pour créer des services utilisant l'IA (chatbot, reconnaissance d'images, etc) qui sont notamment disponibles en cloud.

Historique des révisions du document

Version	Date	Modifications
1.0	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 de 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.

Vous êtes lecteur, expert, fournisseur et avez détecté des erreurs 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 le chrono dans le tableau ci-dessus.

Ce document est téléchargeable à partir de http://www.oezratty.net/wordpress/2018/usages-intelligence-artificielle-ebook.