Opinions Libres
le blog d'Olivier Ezratty

Les GAFA, les entreprises et les données de l'IA

Il est de bon ton de déclarer de manière péremptoire que les GAFA et autres BATX dominent le monde de l'IA et ont accès à un tombereau de données qui leur permettent de l'entraîner et que cela leur assure, ipso-facto, un leadership incontestable dans tous les domaines et tous les métiers, au point de menacer tous les grands acteurs des marchés verticaux.

Au risque d'enfoncer des portes ouvertes, je voudrais contrer cela et expliquer pourquoi cette peur est en grande partie infondée. Elle est notamment liée à une méconnaissance des outils et modèles de l'IA et des données qui les alimentent. Elle relève aussi d'une vision simpliste des applications et de la portée de l'IA d'aujourd'hui et de son incarnation la plus courante, le machine learning.

Ceci n'enlève rien à la puissance des GAFA qui est indiscutable avec leurs plateformes incontournables. Il ne faut cependant pas l'exagérer.

Les données des GAFA et autres leaders

Avec leurs centaines de millions ou milliards d'utilisateurs, les leaders américains de l'Internet collectent de gigantesques volumes de données liées à l'activité de leurs utilisateurs, que leurs services soient gratuits comme pour Google ou Facebook ou payants comme chez Amazon. Ce dont des données issues des usages grand public pour l'essentiel. Les données issues des services gratuits permettent surtout de faire de la publicité ou de la vente ciblée. Même si l'expérience montre que, malgré toutes les beautés du machine learning, ces publicités ciblées sont très "bourrines". Vous avez cherché et acheté un produit en ligne, et hop, vous serez bombardé de publicité "ciblée" sur ce que vous avez déjà acheté. Vous préparez un voyage dans une ville et la publicité ciblée continuera d'agir pendant des mois voir des années alors que vous changez de ville chaque année !

La nature des informations collectée n'est évidemment par la même d'un acteur à l'autre. Ils se complètent pour collecter nos faits et gestes avec dans l'ordre :

- **Google** qui est probablement celui qui en sait le plus sur vous avec vos recherches (de textes, d'images et de vidéos), les sites que vous consultez (via le navigateur Chrome), les vidéos que vous consultez (sur YouTube, voir sur votre TV si elle tourne sous Android), vos déplacements (via Android ou Google Maps/Waze), vos emails (via Gmail), votre agenda voire vos documents de travail (si vous utilisez Google Docs). Il peut aussi connaître votre voix via Google Assistant. Si vous avez l'imprudence d'utiliser ses produits de maison connectée, Google en saura un peu plus sur votre confort thermique et sur votre sécurité.

- **Amazon** qui sait pas mal de choses de vos achats et de vos envie et commence à connaître vos goûts en matière de contenus via vos achats de livres, votre usage de liseuse si vous en avez une ou via Prime Video. Si vous utilisez Amazon Alexa, Amazon vous "écoute" et complète l'ensemble.

- **Apple** qui connaît vos déplacements, les applications que vous utilisez, certains contenus que vous consommez, seulement si vous êtes utilisateur d'iPhone. Comme le modèle économique d'Apple est surtout du vendre du matériel, ils affichent un profil moins invasif dans l'exploitation de vos données, même si l'on découvre parfois qu'ils prennent des libertés.

- **Facebook** connaît aussi vos goûts et sujets de discussion, vos opinions politiques et sait qui sont vos relations et amis. Il peut aussi savoir où vous êtes, notamment en vacances, si vous y diffusez vos photos (dans Facebook comme Instagram).

D'autres leaders du numérique qui ne font pas partie des GAFA au sens littéral du terme en savent aussi pas mal sur vous, mais toujours en pièces détachées :

- **Airbnb** et **booking** connaissent vos voyages, vos envies de voyages, vos habitudes sociales et votre niveau de vie.

- **Uber** connaît une partie de vos déplacements dans la journée et en soirée. Mais pas tous. Pas vos trajets en trains ou en avions, sauf si vous y faites appel partout dans le monde de/vers les gares et aéroports.

- **Netflix** connaît vos gouts en consommation de vidéo et quand vous les consommez. Et donc, quand vous n'en consommez pas… !

- **LinkedIn** – donc Microsoft – connaît une part de votre vie et de votre réseau professionnels.

- **23andme** connaît votre génotype et le commercialise discrètement, anonymisé, à des laboratoires de pharmacie. Peu d'Internautes ont fait appel à ce service en France où il est théoriquement interdit.

Ces gros volumes de données peuvent certainement être croisés mais, en temps normal, ils sont naturellement disséminés sur un grand nombre d'acteurs et de bases de données. Cela ne permet pas de faire n'importe quoi. Le deep learning n'est pas une potion magique qui permet de jouer madame Irma.

A chaque besoin sa solution, son développement logiciel, ses modèles, son entraînement. Les acteurs ci-dessus, surtout les GAFA, ont à leur portée des ressources de data centers immenses permettant d'entraîner de gros modèles de machine learning. Mais ces ressources sont aussi mises à disposition et commercialisées auprès de startups et entreprises. Les GAFA n'en ont donc pas l'exclusive.

La meilleure démonstration de la non-invincibilité des GAFA est l'émergence continue de nouveaux acteurs, souvent américains, qui deviennent les leaders de leur secteur, au nez et à la barbe des GAFA. On a l'habitude de croire que les GAFAMI (en ajoutant Microsoft et IBM) réussissent tout ce qu'ils entreprennent, notamment en termes de diversification. Ce n'est heureusement pas le cas. S'ils ne sont pas les premiers à lancer de nouveaux types de business,

ils se font souvent dépasser par des startups spécialisées comme celles que nous avons citées plus haut (Airbnb, Booking, Uber, …). L'écosystème numérique est ainsi bien moins monolithique et concentré qu'il n'y paraît. Le risque pour les entreprises établies est donc bien plus chez ces acteurs émergents qui s'attaqueraient à leur cœur de métier qu'au niveau des GAFA.

Les données métiers

Malgré l'immensité des données accumulées par les grands acteurs de l'Internet sur nos faits, gestes et envies, ils ne savent pas tout sur vous et n'ont pas accès aux données métiers de nombre de marchés verticaux. En effet, les données métiers des entreprises sont chez elles et uniquement chez elles. Les GAFA n'ont pas mis la main dessus. S'ils le faisaient, cela serait avec l'assentiment des grandes entreprises concernées. A elles de ne pas tomber dans le panneau !

Les exemples sont nombreux. Les données de consommation d'énergie ou d'eau récupérées par les compteurs dits intelligents (qui en pratique, ne le sont pas du tout), les données d'usages de mobiles et du fixe que les opérateurs télécoms possèdent, vos flux et stocks financiers que gèrent vos banques aussi bien dans le grand public que pour les professionnels, vos biens assurés et vos incidents/accidents, etc. Idem pour les retailers qui savent ce que vous achetez chez eux, mais souvent de manière disparate, sauf peut-être pour les supermarchés et hypermarchés si vous utilisez leur carte de fidélité et faites vos courses de manière récurrent au même endroit. Des données qu'ils exploitent d'ailleurs très mal dans des approches marketing personnalisées qui peinent à émerger.

On peut déduire énormément d'informations de la manière dont vous consommez de l'eau ou de l'électricité. La consommation d'eau permet de savoir combien vous êtes dans le foyer, l'âge des occupants et si les gens se lavent. On peut même en déduire certaines pathologies en lien avec la fréquence de visite des WC. Etc ! Cela peut provenir de modèles de machine learning bien entraînés avec des données bien labellisées.

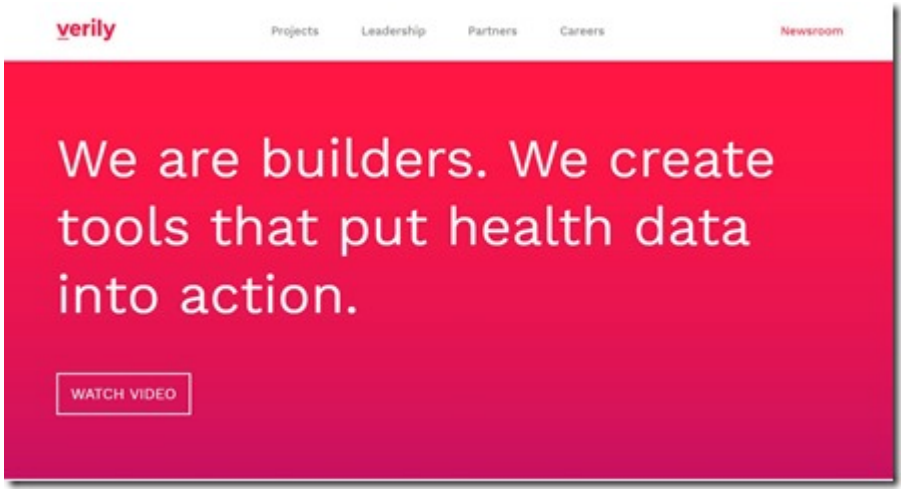
Enfin, n'oublions pas les données que les Etats ont sur nous. Cela dépend évidemment des pays mais ils ont au minimum des données de cadastre, d'état civil, de justice, sur nos véhicules, nos contraventions, nos déclarations fiscales, immobilières et de successions. Cela fait pas mal de données très privées qui ont de la valeur ! La loi Informatique et Liberté de 1978 limite d'ailleurs toujours les capacités de croisement de ces bases de données. La règlementation bancaire protège le secret des données que les banques gèrent.

Enfin, il y a toutes les données générées par les entreprises qui n'ont pas de lien direct avec le grand public comme les industries manufacturières. Elles peuvent évidemment en tirer parti pour améliorer leurs processus de production, de commercialisation et de maintenance.

Et toutes ces données, les GAFA ne les ont pas ! Cela fait un paquet de données ! Et c'est encore moins reluisant pour les BATX, les leaders de l'Internet chinois qui ont beaucoup de données sur les comportements d'Internautes chinois mais, en général, rien de ceux des pays occidentaux. Leurs données ne servent pas à grand chose dans nos marchés aussi bien que les nôtres ou celles des GAFA ne servent à rien pour attaquer le marché chinois.

On commence à s'inquiéter des velléités des GAFA de collecter vos données de santé. Vous pourriez commencer à le faire dans l'application Apple Health sur votre iPhone, en liaison avec votre montre connectée Watch 4 et sa fonction d'électrocardiogramme, mais cela ne va pas très loin. Google aimerait aussi collecter ces données et par la même occasion votre génotype, histoire de les commercialiser de manière directe ou indirecte aux laboratoires de pharmacie pour leurs études épidémiologiques.

En pratique, ces données de santé sont ailleurs. En France, les données de la CPAM comprennent les parcours médicaux et les examens et médicaments prescrits. Les données médicales sont encore réparties dans des bases hospitalières disparates. Et si le DMP (dossier médical personnalisé) commence à poindre du nez, il n'est pas géré par les GAFAs. Il en va de même avec les données du NHS, le système de santé public d'outre-Manche. Aux USA, les bases médicales sont plus nombreuses et connectées dans les hôpitaux et les cliniques privés, mais il n'existe pas de base centralisée du type de celle de la CPAM, sauf pour le cas des programmes Medicaid et Medicair qui concernent les personnes âgées, celles qui sont démunies, les handicapés et les insuffisants rénaux sous dialyse.



Un mythe voudrait que Apple et Google vont immanquablement devenir des "pharma". Les données de santé qu'ils cherchent à accumuler leur donneraient le sésame à ce marché très différent du leur. Ces affirmations témoignent d'une grande méconnaissance du monde des biotechs. Leurs cycles de recherche sont très longs par rapport à ceux du numérique et de l'Internet. Certes, les biotechs font de plus en plus appel à des outils numériques, en particulier pour le drug retargeting et la simulation moléculaire (en IA ou plus tard, avec des algorithmes quantiques). Mais c'est un métier bien à part. Les études de corrélation entre génotype et phénotype souvent mises en avant. Elles ne permettent que de découvrir des facteurs de risques, pas de créer des thérapies ! Et Verily, la filiale santé de Google ? Voir **Que cache Verily, la filiale de Google dédiée aux sciences de la vie ?** de Patrick Randall de Frenchweb qui décrit

bien la situation et conclue que la société se focalise surtout sur la gestion des données et sur les capteurs. Moins sur les thérapies elles-mêmes.

Pour prendre un peu de recul, rappelez-vous ce que les médias et analystes disaient des efforts de Google dans la robotique en 2013 et 2014. A l'époque, leur acquisition de 8 startups dont Boston Dynamics avait fait grand bruit. Ils allaient devenir les rois de la robotique, c'était fichu pour tous les autres acteurs ! Depuis, en quelques années, Google a quasiment abandonné le secteur en cédant Boston Dynamics à Softbank Robotics et en mettant la clé sous la porte d'une autre acquisition de l'époque, Schaft.

Bases de données et intelligence artificielle

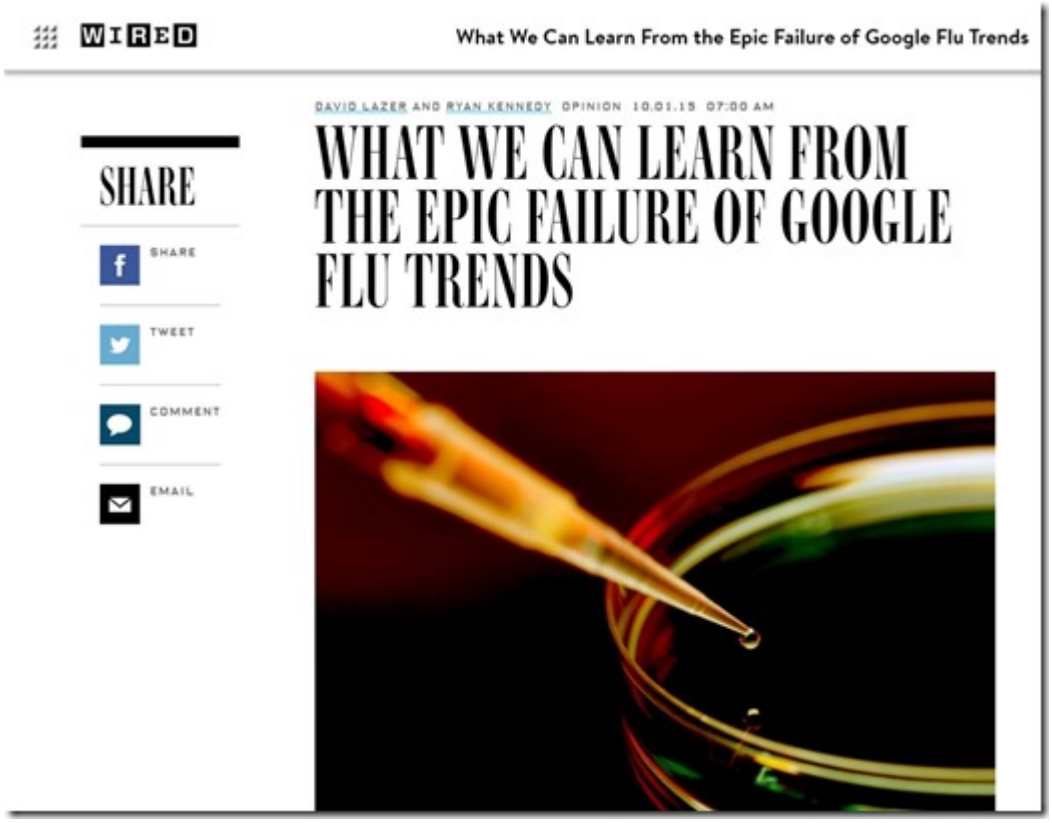
La vision que l'on peut avoir de l'intelligence artificielle est souvent erronée. On imagine un gros système avec plein de données qui est capable de les utiliser de manière omnipotente et omnisciente d'un coup de click. Bref, l'entrepôt de données universel qui sert à tout prédire. L'oracle ultime.

La mise en œuvre de l'IA dans le monde réel est bien plus ingrate et laborieuse. On entraîne des modèles de machine learning au cas par cas. Ces modèles permettent de labelliser des situations (segment client), des images (les objets qu'elles contiennent), des émotions (via de la biométrie et du texte) et de faire diverses prédictions. L'entraînement des modèles de machine learning exploite généralement des données anonymisées. Chaque fonction demande un modèle. Leur entraînement prend du temps et est réalisé pour chaque besoin spécifique. Le nettoyage et la préparation des données à exploiter est un énorme travail.

Ce modèles adoptent une vision probabiliste des problèmes à résoudre. Ils ont besoin d'échantillons importants pour bien fonctionner. Ils prédisent des valeurs futures en fonction des données du passé. Si le passé n'est pas satisfaisant pour une raison ou une autre, les modèles de machine learning vont reproduire bêtement ces insatisfactions.

On peut aussi créer des modèles de machine learning en croisant des données disparates. Cela permet éventuellement de découvrir des corrélations entre données, mais pas forcément les explications associées. Il faut toujours compléter cette approche par du bon sens métier.

Les données qui servent à créer des modèles de machine learning sont principalement de quatre formats différents : données structurées et tabulaires (bases de données SQL, tableaux, …), d'images et vidéo, audio et textuelles. Les GAFA maîtrisent bien les données ouvertes indexables sur Internet et celles qui sont fournies explicitement ou implicitement par les utilisateurs.



Mais cela ne fonctionne pas à tous les coups. Comme le rappelle bien **What we can learn from the epic failure of Google Flu Trends** de David Lazer et Ryan Kennedy en octobre 2015, Google avait voulu identifier et cartographier l'émergence d'épidémies de grippes avec Google Flu Trends entre 2008 et 2015. Google pensait pouvoir prédire ces épidémies plus rapidement que le CDC (Center for Diseases Control US) via les recherches en ligne sur le sujet. Mais il rata complètement la détection de la saison 2013. L'article pointe le besoin de croiser des données et de les rendre ouvertes. En clair, de rendre celles de Google accessibles au CDC !

L'autre croyance répandue est la confusion entre outil de développement et modèle. Ainsi, certains peuvent rapidement croire qu'en utilisant les bibliothèques ou les outils de TensorFlow, qui proviennent de Google, ils vont aider Google à entraîner ses propres IA. C'est une vision naïve des choses qui relève de la difficulté à séparer code et données associées, comme dans un logiciel traditionnel. Si vous créez un modèle avec TensorFlow et l'entraînez avec vos données, cela va générer un modèle entraîné pour vos propres besoins et Google n'en aura que cure. Par contre, vous bénéficiez du travail d'une grande communauté de développeurs qui ont créé des modèles prêts à l'emploi pour ce framework, une part d'entre eux provenant de Google.

On oublie aussi le lien indissociable entre les données des GAFA et les caractéristiques de leurs services et à leur valeur. Ces données ne tombent pas du ciel. Elles sont directement liées aux services associées qui permettent de les collecter. Facebook collecte les données que vous lui fournissez en racontant votre vie et en *likant* des pages. Google récupère les liens que vous créez entre les sites web pour les indexer et gérer leur pagerank. Etc. Il est rare que l'on puisse découpler les données et les services associés. Donc, avant de prêter aux GAFA toutes les innovations du monde dans l'IA, posez-vous les questions des services que vous pourriez créer avec vos propres données, éventuellement croisées avec des données tierces.

Autre méconnaissance, le volume des données nécessaires à l'entraînement d'un modèle de machine learning. Celui-ci est très variable selon les besoins, les données, leur variabilité et les applications. Pour optimiser l'entraînement des modèles, même les GAFA font le tri dans les

données et ne les exploitent pas intégralement. Elles seraient trop longues à ingérer pour l'entraînement des modèles de machine learning.

Les entreprises de secteurs traditionnels peuvent toujours se plaindre du monopole qu'ont acquis les GAFA pour toucher les Internautes. Ils disposent en effet des moyens de cibler les audiences grand public avec une assez bonne précision. Mais ce travail d'intermédiation a toujours existé d'une manière ou d'une autre. C'est le marché de la publicité qui a évolué. Il est passé d'un mix de médias à sens unique et à faibles interactions avec les consommateurs, à des médias interactifs qui facilitent ce fameux ciblage. C'est un métier et une spécialité.

Enfin, du côté de la préservation de la vie privée, il existe un nouveau champ scientifique appelé *differential privacy* qu'il faut prendre en compte pour éviter qu'un modèle de machine learning entraîné permette l'identification d'individus selon des critères précis. Exprimé simplement, ce procédé consiste à ajouter du bruit dans les données et dans les modèles lors de l'entraînement et lors d'inférences pour éviter cette identification, en particulier lorsque l'on cherche à faire une prédiction sur un échantillon réduit d'individus.

Conséquences

La principale leçon de ces observations est que les GAFA n'empêchent pas vraiment les grandes entreprises françaises d'adopter le machine learning et d'exploiter leurs propres données pour ce faire. Il en va de même pour les PMEs de nombreux secteurs d'activité. Les blocages sont autres : la capacité à faire de la veille technologique, à identifier des besoins mal couverts par l'existant, à innover, à sortir des sentiers battus et à trouver et/ou former les bonnes compétences.

La seconde est bien connue : analysez bien vos données ! Votre entreprise en possède beaucoup plus qu'elle ne le croit. Elles proviennent de capteurs, d'usines, des outils de la relation client, sites web et applications mobiles. De gros volumes de données faiblement structurées issues des outils de communication interne et externe, et collaboratifs. sont aussi exploitables par les nouveaux outils de traitement du langage. Il faut aussi trouver des idées de croisement de données métiers, notamment avec des données ouvertes. Enfin, il faut évidemment gérer des données "vivantes", fréquemment mises à jour, de qualité et bien labellisées. Cela requiert un processus qualité ! Il faut aussi bien analyser la structure statistique et sémantique des données pour identifier leurs biais éventuels. L'expertise en machine learning consiste à avoir une idée de la volumétrie des données nécessaire pour entraîner un modèle. Il y a un savoir partageable et… les tests !

La troisième est de recentrer le débat sur la nature des services à créer qui pourraient exploiter toutes ces données. Se lancer bille en tête dans la création d'un entrepôt de données pour le machine learning (dit "datalake") sans avoir la moindre idée des services pouvant l'exploiter n'est pas toujours la voie du succès. Au passage, en n'oubliant pas que le machine learning n'est qu'une des branches techniques de l'IA. Il en existe d'autres qui sont moins voraces en données comme les moteurs de règles, les ontologies, les arbres de décisions et autres graphes ou les systèmes multi-agents. En particulier, la robotique fait beaucoup appel à ces techniques.

La quatrième est : n'exposez pas forcément vos données à tous vents. C'est l'épée à double tranchant de l'open data : c'est bien au premier abord, mais les GAFA peuvent les exploiter et plateformiser votre activité. C'est le paradoxe de l'ouverture, notamment pour les Etats qui ouvrent les données. D'où la question de l'usage de ces données à but lucratif et la consolidation de monopoles associés. A contrario, creusez l'opportunité de mutualiser vos données avec

d’autres acteurs d’un même marché ou de marchés adjacents pour créer plus de valeur. L’union peut faire la force face aux GAFA.

Bref, ne vous endormez pas sur vos lauriers ou dans votre pré-carré !

Cet article a été publié le 14 juillet 2019 et édité en PDF le 15 juillet 2019.