Les avancées de l'intelligence artificielle

Mars-Mai 2016

Olivier Ezratty

A propos de l'auteur

Olivier Ezratty olivier (at) oezratty.net , http://www.oezratty.net , @olivez

Conseil en Stratégies de l'Innovation +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 métiers de l'image dans les médias numériques (TV, cinéma, photographie). Il leur apporte une triple expertise : 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**, une association membre d'Initiative France qui accompagne et finance l'amorçage de startups franciliennes.
- Membre depuis fin 2015 du Comité de Prospective de l'**ARCEP**.
- Membre du jury de divers **concours entrepreneuriaux** comme le Grand Prix de l'Innovation de la Ville de Paris, de Systematic et la Startup Academy.
- Expert auprès du pôle de compétitivité **Cap Digital** ainsi que de la **Caisse des Dépôts** et du **CNC**.
- Mentor dans de nombreux **Startups Weekends**, notamment à Nantes, Rennes, Sophia-Antipolis, Bordeaux, Montpellier et à l'Ecole Polytechnique.
- Advisor, board member et/ou consultant dans quelques startups comme **Voluntis**.

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 150 000 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) qui vise à augmenter la place des femmes dans les métiers du numérique.

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 : le robot Pepper d'Aldebaran Robotics, prise par l'auteur à Tokyo en octobre 2014.

Table des matières

Introduction

Cet ebook est une compilation de neuf articles sur les Avancées de l'Intelligence Artificielle publiés entre mars et mai 2016 sur le blog Opinions Libres.

Ces articles avaient comme ambition de décortiquer l'état de l'art de l'intelligence artificielle en grattant derrière les effets d'annonces, notamment autour de la victoire de Deep Mind contre le champion du monde du jeu de Go.

Ils ne prétendent pas remplacer un ouvrage de référence sur l'intelligence artificielle. Il y en a plein, notamment dans l'édition américaine. Certains de ces ouvrages sont évoqués dans ce document. Nombreux sont ceux qui racontent à leur manière l'histoire de l'intelligence artificielle, avec ses hauts et ses bas, depuis sa naissance pendant les années 1950. Nous sommes actuellement dans une période « haute ».

Mes articles ont évolué après leur publication, comme des sables mouvants. Notamment ceux qui concernent les startups de l'intelligence artificielle que j'ai pu compléter au fil de l'eau et des sollicitations. Au gré de mes lectures, j'ai complété et affiné le texte.

Voici la synthèse et la structure de ce document :

- **Sémantique et questions clés** : qu'est-ce que l'IA ? Comment est-elle segmentée ? Quelles sont les grandes questions qui se posent à son sujet ?
- **Histoire et technologies de l'intelligence artificielle** : comment l'IA a-t-elle progressé depuis les années 1950 ? Quelles sont ses principales briques technologiques, surtout algorithmiques ? Les progrès récents viennent-ils du logiciel, du matériel ou des données ?
- **IBM Watson et le marketing de l'intelligence artificielle** : qu'est-ce qui se cache derrière IBM Watson ? Comment interpréter sa performance dans sa victoire au Jeopardy en 2011 ? Quelles sont ses autres applications, notamment dans la santé ? Remplace-t-il les experts ? Quelle est la stratégie d'IBM ? Comment les startups de l'IA s'y prennent-elles dans leur marketing ?
- **Les startups US de l'intelligence artificielle** : quelles sont les principales startups américaines de l'IA ? Lesquelles sont les mieux financées ? Quelles plateformes ont le vent en poupe ?
- **Les startups acquises par les grands du numérique** : comment les grands acteurs américains du numérique investissent-ils dans l'IA et quelles startups ont-ils acquis ?
- **Les startups françaises de l'intelligence artificielle** : comment fonctionne la recherche française dans l'IA ? Quelles sont les principales startups du secteur ? Où sont les opportunités ? Quelle stratégie bâtir dans l'IA ?

- **La modélisation et la copie du cerveau** : quelle est notre compréhension actuelle de la structure du cerveau ? Est-ce que les prévisions de la singularité visant à copier le contenu d'un cerveau dans un ordinateur sont réalistes ?
- **Les évolutions de la loi de Moore** : comment évolue la loi de Moore dans la pratique, à la fois dans les supercalculateurs et dans les produits grand public ? Comment va évoluer le logiciel et le matériel pour faire évoluer les applications à l'intelligence artificielle ?
- **La robotisation en marche des métiers** : 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 ? Qu'en est-il de l'IA dans la politique ?
- **Epilogue** : synthèse d'impressions et que penser de tout cela ?

Bonne lecture !

Sémantique et questions clés

L'intelligence artificielle fait partie de ces technologies qui peuvent générer toutes sortes de fantasmes pour les uns et de craintes pour les autres. Peut-être bien plus qu'avec la majeure partie des technologies, peut-être à l'exception de la génomique, il nous est difficile d'en comprendre le fonctionnement, l'état de l'art et d'en apprécier les enjeux associés.

L'actualité technologique nous ressasse des performances d'**IBM Watson**, des robots quadrupèdes ou bipèdes de **Boston Dynamics**, la future ex-filiale de Google, ou des dernières expérimentations de **voitures à conduite automatique**, chez Google ou ailleurs. Il est facile de mettre tout cela dans le même sac, sous l'appellation d'intelligence artificielle comme si c'était un tout bien unifié. Il n'en est rien !

L'intelligence artificielle est un pan entier de l'informatique avec sa diversité, ses briques technologiques, ses assemblages et solutions en tout genre. Qui plus est, elle est aussi liée à d'autres sciences : les mathématiques, les statistiques, les sciences cognitives, la psychologie et la neuro-biologie.

C'est un véritable écosystème hétéroclite. Qui plus est, la grande majorité des solutions commerciales d'Intelligence Artificielle sont faites de bric et de broc, en fonction de besoins très spécifiques. On est loin d'avoir sous la main des solutions d'IA génériques.

Les craintes sur l'intelligence artificielle

Cette diversité bio-informatique génère pour l'instant une sorte de protection contre les menaces de l'IA évoquées par certains Cassandre de l'industrie. Mais jusqu'à quand ?

Les Cassandre redoutent une échéance fatidique où l'IA prendra le pas sur l'intelligence humaine. Le célèbre 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 !

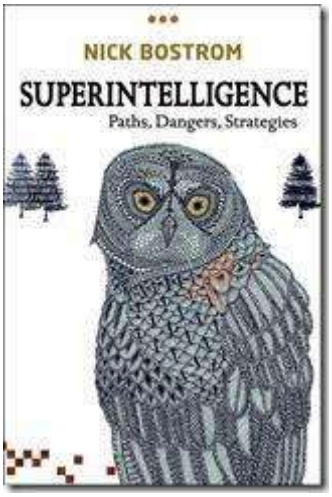
Il reprenait à son compte, en version pessimiste, une citation de Irwin John Good datant de 1965 publiée dans **Speculations Concerning the First Ultraintelligent Machine** selon laquelle la machine ultra-intelligente sera la dernière invention que l'homme aura besoin de créer.

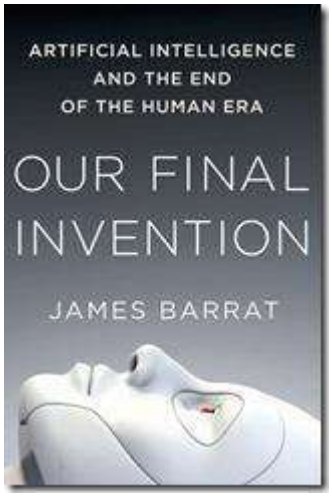
9. Conclusions

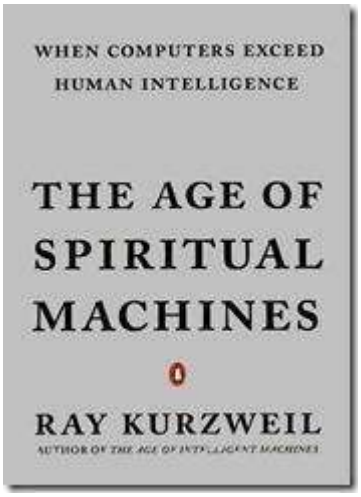
These "conclusions" are primarily the opinions of the writer, as they must be in a paper on ultraintelligent machines written at the present time. *In the writer's opinion then:*

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 nombreuses ouvrages, comme celles 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.







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 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, 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[1].

Bill Joy était en fait effrayé des perspectives avancées par Ray Kurzweil qu'il avait rencontré dans une conférence en 1998 et 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 de-

[1] On y apprend d'ailleurs qu'il avait rencontré Jacques Attali et que ce dernier avait indirectement influé le cours des événements de Java !

mander une pause technologique et une réflexion sur les limites à ne pas dépasser avec l'intelligence artificielle comme avec la robotique.

Il existe même des instituts de recherche qui planchent sur cette question : le **Center for the Study of Existential Risk** de Cambridge et le **Future of Humanity Institute** d'Oxford.

Dans les optimistes, semble-t-il moins nombreux, on trouve bien évidemment le pape actuel de la singularité **Ray Kurzweil** qui anticipe celle-ci autour de 2030-2040 ainsi que **Mark Zuckerberg** de Facebook qui pense que l'homme sera raisonnable dans ses usages de l'IA.

Les optimistes sont aussi souvent les spécialistes de l'IA qui voient de près l'ingratitude de la discipline et estiment en général que l'on est loin 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 !

Grandes questions sur l'intelligence artificielle

Seulement voilà, l'intelligence artificielle n'est pas un produit. Ce n'est pas non plus un logiciel unifié comme un traitement de texte, une application mobile ou même un système d'exploitation.

Il n'y a pas de logiciel d'intelligence artificielle mais *des* solutions d'intelligence artificielle qui s'appuient sur des dizaines de briques différentes qui vont de la captation des sens, notamment audio et visuels, à l'interprétation des informations, au traitement du langage et à l'exploitation de grandes bases de données structurées ou non structurées. Leur intégration reste encore une affaire de bricolage. Nous en sommes toujours à l'âge de pierre, avec seulement une cinquantaine d'années de recul sur la question.

Dans la lignée d'autres séries de défrichages de sujets technologiques complexes publiées sur « Opinions Libres », je me propose ici de décortiquer ce que l'on sait de l'état de l'art de l'Intelligence Artificielle, de ses applications et des progrès en cours.

Je vais notamment tenter de répondre à plusieurs questions clés qui me travaillent :

- Quelles sont les **grandes briques technologiques** de l'intelligence artificielle ? C'est un domaine un peu fouillis que je vais essayer de segmenter. Sans forcément adopter le découpage des ouvrages de référence sur le sujet qui sont assez difficiles d'abord.
- Quels sont les différents **usages de l'intelligence artificielle** ? Je vais reprendre les études de cas les plus courantes et les commenter.
- Comment les solutions d'intelligence artificielle sont-elles **commercialisées**, en prenant l'exemple d'IBM Watson ? En décrivant l'approche qui est actuellement à dominante service pour les solutions d'entreprises, mais avec un fort développement d'applications grand public en parallèle. Comment l'IA se retrouve-t-elle dans le marketing des startups ?
- Comment se développe **l'écosystème de l'intelligence artificielle**, des grands groupes comme Google, Facebook, Microsoft et IBM jusqu'aux start-ups du secteur ? Quels sont les enjeux industriels dans le secteur ? Et la position de la France ? Quel est le rôle de l'open source ?
- Comment les **briques d'intelligence artificielle progressent-elles** ? Est-ce lié à l'invention de nouveaux procédés techniques, aux progrès du matériel ou aux deux, et dans quelle proportion ? Qu'est-ce qui pourrait accélérer ou ralentir ces progrès ?
- Quel sera l'**impact de l'IA sur le futur de l'emploi** ? Est-ce réaliste de prévoir que la moitié des emplois disparaitront dans à peine deux décennies ? Quels sont les emplois les plus menacés ? Quels sont les moyens d'éviter de se faire « robotiser » ? Quid des politiques et du fonctionnement des états au passage ?

Il s'agit ici du résultat d'une quête personnelle sur un sujet nouveau, s'appuyant en grande partie sur une recherche bibliographique extensive. Je ne suis pas spécialiste de ce domaine et j'apprends au fil de l'eau tout en partageant le résultat de cet apprentissage.

Qu'est-ce que l'intelligence artificielle ?

L'Intelligence Artificielle regroupe les sciences et technologies qui permettent d'imiter, d'étendre et/ou d'augmenter l'intelligence humaine avec des machines.

L'IA (AI en anglais) a été conceptualisée en 1956 par John McCarthy, Alan Newell, Arthur Samuel, Herbert Simon et Marvin Minsky, ce dernier étant décédé en janvier 2016.

Cela s'appuyait comme toute innovation, progrès scientifique ou nouvelle théorie sur de nombreux travaux et visions antérieurs à 1956 : le concept de calculus ratiotinator de Leibnitz, la machine et le test de Turing, les neurones formels de McCullochs et Pitts, l'architecture de Von Neuman, le théorème de Shanon, etc.

L'IA s'inscrit dans une longue tradition humaine d'innovations s'appuyant d'abord sur la force mécanique puis sur la force intellectuelle toutes deux artificielles. Certains scientifiques visent à atteindre dans un premier temps l'intelligence humaine.

Mécaniquement, les effets de levier technologiques sont tels qu'un seuil aboutirait au dépassement rapide de l'intelligence humaine par celle de la machine.

L'IA fait partie de ce que l'on appelle aussi les sciences cognitives. IBM intègre ainsi Watson dans son offre de "cognitive computing". J'ai cherché comment on pouvait la segmenter en domaines. En gros, on trouve d'abord ce qui concerne 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.

J'ai tenté ensuite de segmenter le domaine de l'IA et bien mal m'en a pris. Plusieurs découpages existent, au niveau conceptuel puis au niveau technique.

De l'IA faible à l'IA forte

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. Cela a commencé avec les systèmes jouant et gagnant aux échecs comme Deep Blue d'IBM en 1997, puis avec des systèmes experts pointus comme dans certains secteurs de la santé.

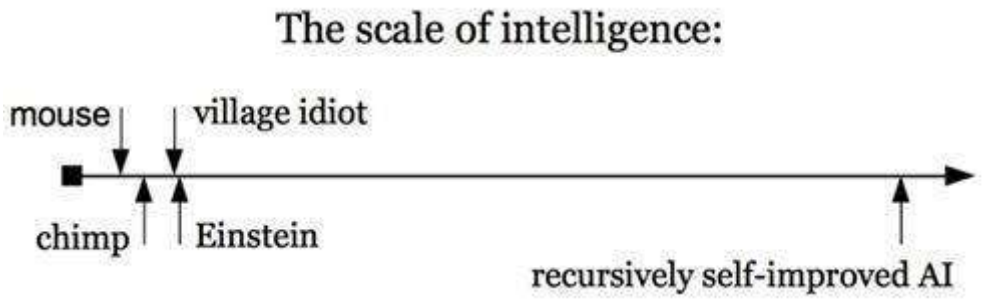
On peut y mettre en vrac les moteurs de recherche courants, la détection de fraudes bancaires, le credit rating de particuliers, la conduite automatique ou assistée, Apple SIRI, Microsoft Cortana et Google Now. Si l'IA n'imite pour l'instant pas toujours l'homme, la force brute et l'usage d'éléments 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 dans d'autres dimensions que celles qui font que homme est l'homme. Par contre, ne font pas partie du champ de l'IA les problèmes simples qui peuvent être résolus avec de simples algorithmes. C'est le cas des systèmes de pilotage automatiques d'avions.

L'**Artificial General Intelligence** (AGI) correspond 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. On peut intégrer dans ce niveau un grand nombre des capacités humaines : l'usage du langage à la fois comme émetteur et récepteur, l'usage de la vue et les autres sens, la mémoire et en particulier la mémoire associative, la pensée, le jugement et la prise de décisions, la résolution de problèmes multi-facettes, l'apprentissage par la lecture ou l'expérience, 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.

On peut 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

en est encore loin, même si certaines de ces capacités notamment linguistiques et de raisonnement général sont en train de voir le jour.

L'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).



L'**Artificial Super Intelligence** (ASI) est la continuité logique de l'étape précédente, liée à la puissance des machines qui se démultiplie et se distribue plus facilement que celle d'un cerveau humain avec ses entrées-sorties et ses capacités de stockages et de traitement limitées. Qui plus est, cette intelligence peut 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 ce niveau, l'intelligence de la machine dépasse 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.

Il l'est également dans cet essai **The Singularity – A philosophical analysis** du philosophe australien David J. Chalmers qui propose notamment de tester d'abord l'ASI dans un environnement entièrement virtuel entièrement déconnecté du monde réel pour tester ses aptitudes. Si cela peut rassurer [2]!

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.

[2] 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.

Toutes ces conjectures semblent 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.

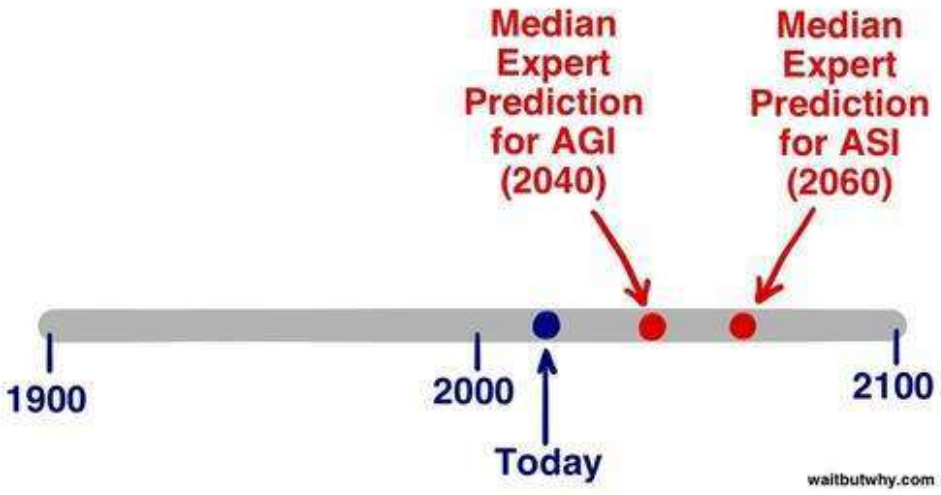
Dans la pratique, l'IA d'aujourd'hui va déjà bien au-delà des capacités humaines, notamment lorsque la mémoire est en jeu. La capacité des systèmes experts, et notamment d'IBM Watson, à brasser d'énormes volumes d'information fournit des capacités inaccessibles à n'importe quel humain, même surdoué. 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 deviendra.

L'IA permettrait d'atteindre l'immortalité

Vu du versant de l'optimisme, 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.

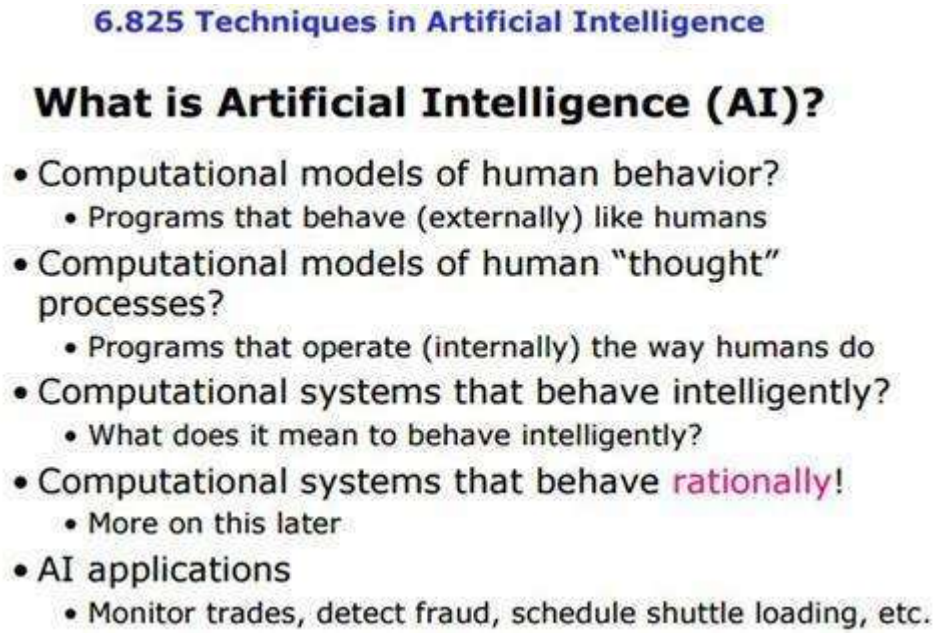
D'ailleurs, les meilleures solutions d'IA comme l'usage d'IBM Watson dans la cancérologie s'appuient sur le corpus issu de toutes ces expérimentations. Il a une base physique et réelle. On pourra certainement automatiser l'expérimentation biologique encore plus qu'aujourd'hui dans la recherche de thérapeutiques, 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.

Définitions et segmentations de l'intelligence artificielle

Poursuivons notre quête de la définition de l'IA dans un **cours du MIT**. L'IA serait un ensemble de techniques permettant d'imiter le comportement humain, agissant de manière rationnelle en fonction de faits et données et capables d'atteindre un objectif. 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 et notre capacité d'optimisation.

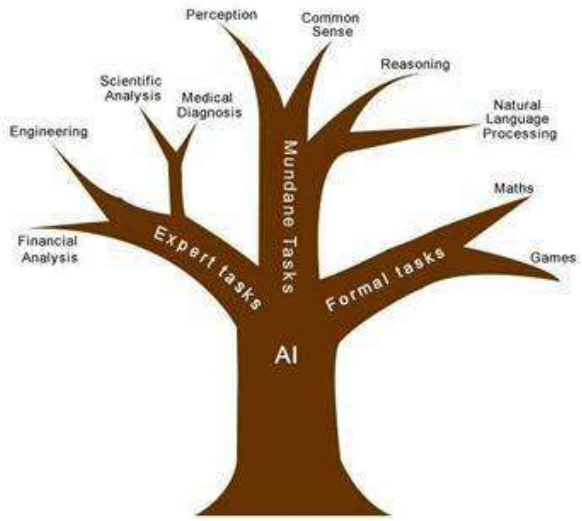


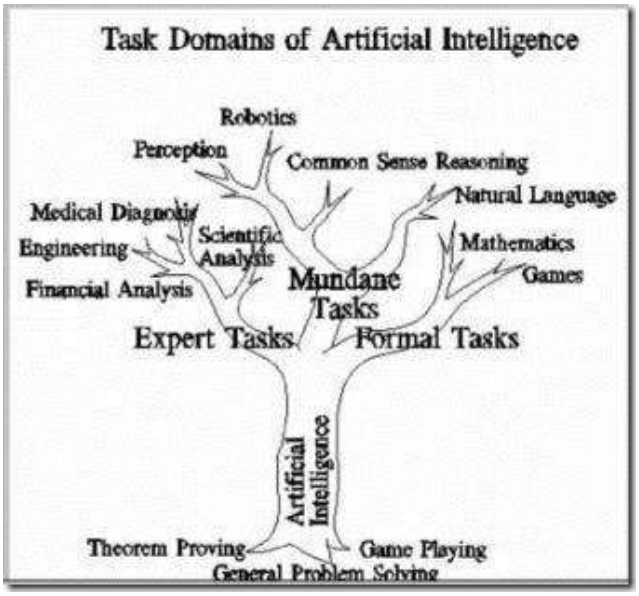
Autre découpage, plus fin, de l'IA, en trois domaines :

- 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.
- Le **connectionisme** 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.
- 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 d'Aldebaran Robotics / Softbank.

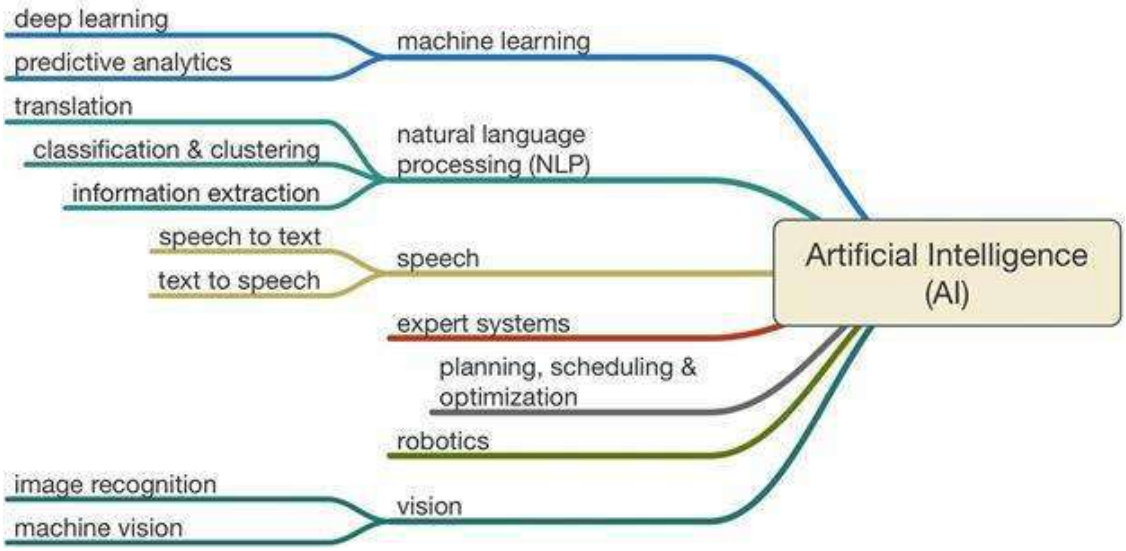
L'IA peut notamment servir à automatiser des **processus cognitifs** et en s'appuyant sur quatre étapes : l'observation des faits et données, leur interprétation, leur évaluation et la décision, avec une action ou une proposition d'actions, souvent basée sur des statistiques.

Voici encore un autre découpage : celui de l'arbre avec 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). Mais cela ne décrit pas les briques technologiques associées pour autant.



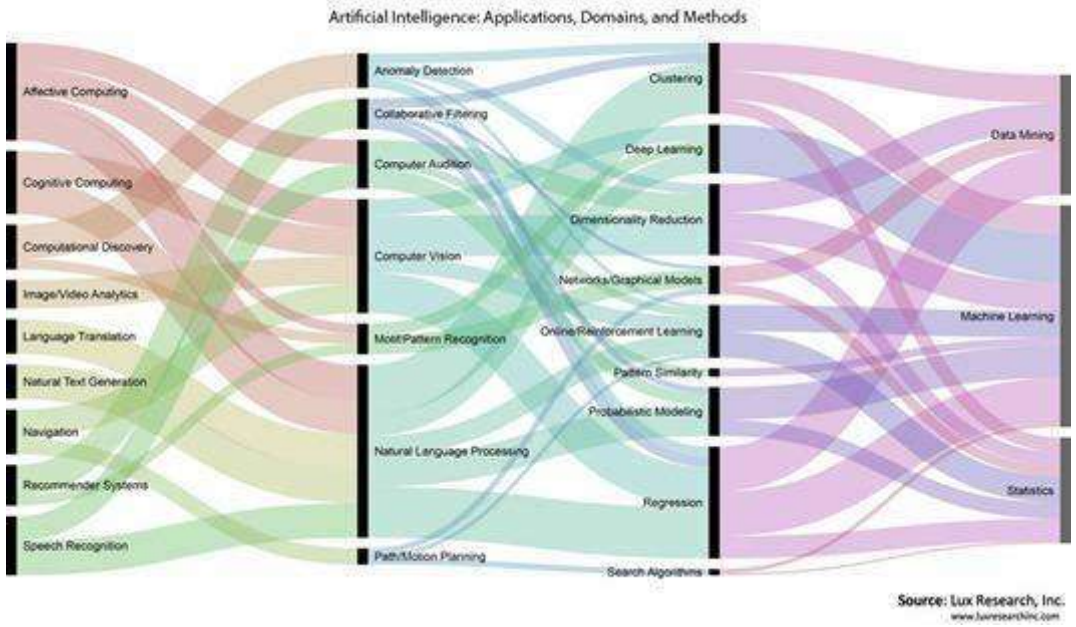


Cet autre découpage, plus terre à terre, comprend le machine learning, le traitement du langage, les systèmes experts, la robotique et la vision. L'architecture est moyenne : il serait plus logique de regrouper les sens avec la parole et la vision. Sans compter l'ouïe qui peut aussi servir. Quand à la robotique, elle a vocation à intégrer tous les autres champs du schéma et à en ajouter d'autres qui lui sont spécifiques comme ceux des capteurs, des matériaux, de la mécanique, des moteurs électriques et des batteries.



Enfin, ce **dernier schéma** fait un lien formel entre trois groupes :

- Les **applications** (affective computing, reconnaissance d'images et vidéos, traduction, …).
- Les **domaines d'applications** (computer vision, NLP, …).
- Les **méthodes** (avec trois grandes catégories : le data mining, le machine learning et les statistiques). Bien bien, mais on peut aussi faire du data mining grâce à du machine learning et ce dernier peut aussi s'appuyer sur des statistiques. Tout cela est bien récursif !



Cela rappelle à bon escient que les solutions à base d'IA sont des assemblages de diverses briques logicielles selon les besoins. Et 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.

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, notamment dans la reconnaissance de la parole.

Si on demandait à un système d'Intelligence Artificielle de s'auto-définir et s'auto-segmenter en exploitant les données bibliographiques disponibles, il serait bien mal en point ! Un peu comme il est difficile de caractériser une période contemporaine, sans le regard de l'historien du futur qui pourra prendre du recul pour analyser le présent.

Vous êtes déjà paumés ? Moi aussi ! Ce qui n'empêche pas de continuer ce parcours !

Histoire et technologies de l'intelligence artificielle

Passons à un côté plus terre à terre en faisant un petit inventaire approximatif des techniques de l'IA. Il s'agit toujours de vulgarisation et d'une restitution de mon processus de découverte du sujet !

Nous évoquerons en partie la question du matériel, notamment pour les réseaux de neurones. Le reste le sera dans un chapitre dédié aux évolutions de la loi de Moore.

Des hivers au printemps de l'IA

L'histoire moderne de l'intelligence artificielle a démarré en 1957. S'en est suivi une période de recherche fondamentale importante, notamment au **MIT AI Lab**, à l'origine notamment du langage **LISP**, créé en 1958, qui servit pendant deux à trois décennies à développer des solutions logicielles d'IA.

Ces recherches étaient principalement financées par l'ARPA, l'agence de recherche du Pentagone devenue ensuite la DARPA, l'équivalent de la DGA française, mais évidemment bien mieux financée.

La recherche sur l'IA était principalement 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. Ce qui peut alimenter au passage les craintes sur les applications futures de l'IA, notamment lorsqu'elle atteindra le stade de l'AGI (IA généraliste).

L'IA connu son premier "hiver" avec une réduction d'une bonne part de ces budgets à partir de 1973, tant au Royaume-Uni qu'aux USA. C'était la conséquence de la publication du **Rapport Lighthill** destiné à l'organisme public britannique **Science Research Council** – équivalent de notre Agence Nationale de la Recherche française – 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 n'étaient pas encore bien développées à cette époque. C'est un bel exemple de manque de vision long terme des auteurs de ce rapport.

Cet hiver a duré jusqu'en 1980. A noter que la période 1973-1980 correspond au premier âge de la micro-informatique, avec la création de Microsoft (1975), d'Apple II (1977) puis les préparatifs du lancement de l'IBM PC (1980-1981).

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.

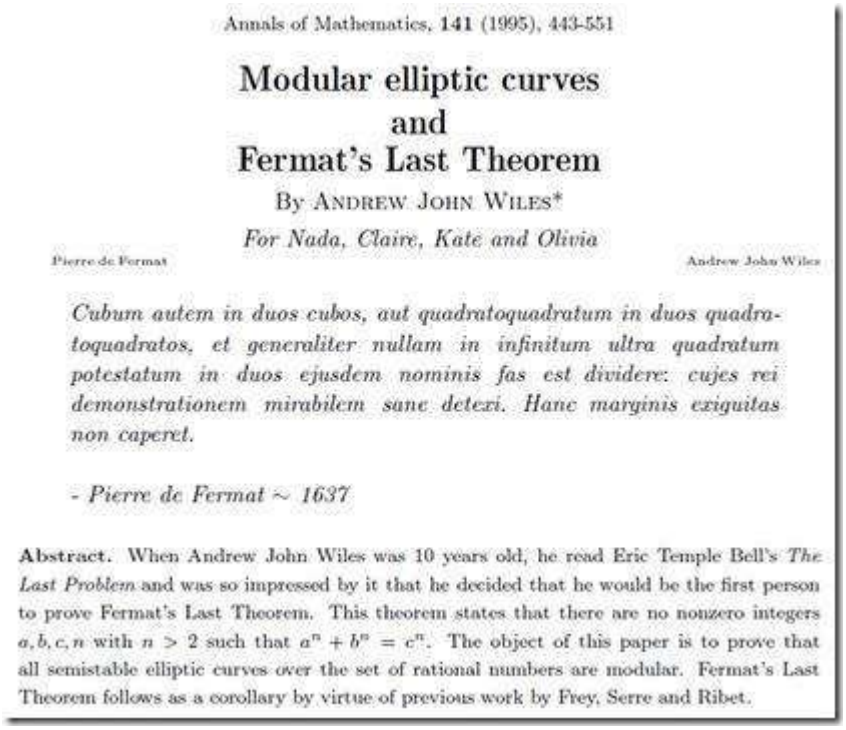
Cette **histoire de l'IA** en fait un inventaire intéressant. **Herbert Simon** (*ci-dessus*) et **Allen Newell** prévoyaient en 1958 qu'en 10 ans, un ordinateur deviendrait champion du monde d'échecs et qu'un autre serait capable de prouver un nouveau et important théorème mathématique. 30 ans d'erreur de timing pour la première prévision et autant pour la seconde sachant qu'elle est toujours largement en devenir pour être générique !

Cet écueil est le même dans les prévisions actuelles autour de la singularité et du transhumanisme : l'ordinateur plus intelligent que l'homme en 2030 ou 2040 et l'immortalité pour les enfants qui viennent de naitre !

Le chercheur d'IBM **Herbert Gelernter** avait réussi en 1958 à utiliser un logiciel de démonstration de théorèmes de géométrie fonctionnant en chainage arrière - de la solution jusqu'au problème - sur un IBM 704 et à partir d'une base de 1000 règles. Cela relevait d'une combinatoire plutôt simple. C'était prometteur.

Il en va autrement 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 impossible pour un entier n>2) qui n'ont jamais été démontrés via de l'IA.

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. 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 !



Annals of Mathematics, 141 (1995), 443-551

Modular elliptic curves
and
Fermat's Last Theorem

By ANDREW JOHN WILES*

For Nada, Claire, Kate and Olivia

Pierre de Fermat — Andrew John Wiles

Cubum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos, et generaliter nullam in infinitum ultra quadratum potestatum in duos ejusdem nominis fas est dividere: cujes rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

- Pierre de Fermat ~ 1637

Abstract. When Andrew John Wiles was 10 years old, he read Eric Temple Bell's *The Last Problem* and was so impressed by it that he decided that he would be the first person to prove Fermat's Last Theorem. This theorem states that there are no nonzero integers a, b, c, n with $n > 2$ such that $a^n + b^n = c^n$. The object of this paper is to prove that all semistable elliptic curves over the set of rational numbers are modular. Fermat's Last Theorem follows as a corollary by virtue of previous work by Frey, Serre and Ribet.

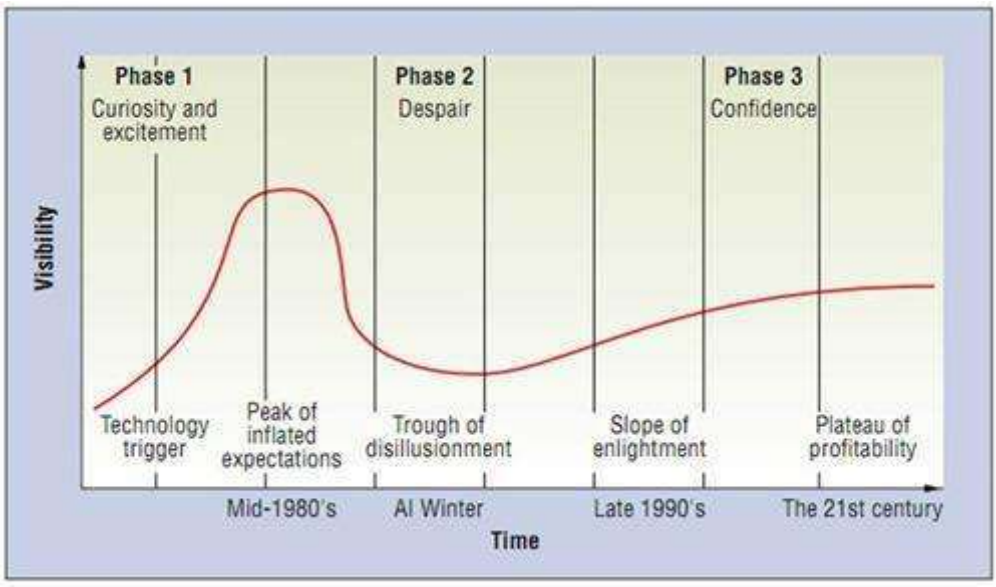
Herbert Simon prévoyait aussi – toujours en 1958 – qu'en 1978, les machines seraient capables de réaliser toutes les activités intellectuelles humaines. Et la loi de Moore n'existait pas encore puisqu'elle a été énoncée après cette prévision, en 1965 !

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 ?

Les retards étaient aussi 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 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 !

Après ce premier hiver de l'IA, s'en est suivie une période d'enthousiasme au début des années 1980 alimentée notamment par la vague des systèmes experts. Le langage **Prolog** du français Alain Colmerauer a contribué à cette vague.

Une nouvelle vague de désillusions s'en est suivie autour des années 1990. L'une des raisons 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.



(**source du schéma** ci-dessous)

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** lancé à £350 million et enfin le **Strategic Computing Initiative** de la DARPA. Tous ces projets ont capoté et ont été terminés 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, linéaire et centralisée.

La fin des années 1980 a aussi connu l'effondrement du marché des **ordinateurs dédiés au langage LISP**.

Pendant les années 1990 et 2000 ont émergé de nombreux projets de **HPC** (high-performance computing), 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.

Depuis le début des années 2000, l'IA a été relancée grâce à diverses évolutions :

- 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.
- 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 au jeu de Go contre son champion du monde.

- L'**Internet** 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.
- La disponibilité de très **gros volumes de données**, via les usages de l'Internet et des mobiles, des objets connectés ou de la génomique, qui permet d'associer les méthodes de force brute et les réseaux neuronaux et autres machine learning ou méthodes statistiques.
- Les **besoins** dans la robotique, dans la conquête spatiale (Curiosity, Philae…), dans les véhicules à conduite assistée ou autonome, dans la sécurité informatique, ainsi que dans la lutte contre la fraude et les scams et pour pourvoir à ces besoins toujours aussi prégnants de s'occuper des personnes âgées au Japon.
- 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 et le e-commerce.
- L'adoption de **méthodes** scientifiques et pragmatiques – basées sur l'expérimentation – et transdisciplinaires, par les chercheurs et industriels.

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 s'opposer les partisans du connexionnisme – utilisant le principe des réseaux de neurones et de l'auto-apprentissage – face à ceux du computationnisme qui préfèrent utiliser des concepts de plus haut niveau sans chercher à les résoudre via des procédés de biomimétisme.

On retrouve cette dichotomie dans la **bataille entre "neats" et "scuffies"**, 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. A ce jour, les méthodes les plus couramment utilisées sont plutôt des domaines mathématiques et procéduraux, mais les méthodes à base de réseaux neuronaux et d'auto-apprentissage font leur chemin. L'intelligence artificielle intégrative qui se développe vise à exploiter conjointement toutes les approches.

Aujourd'hui, l'IA est aussi l'objet d'un débat de société, philosophique, économique (sur le futur de l'emploi) et donc politique. Les débats ont tendance à trop sortir de la sphère scientifique et technique, au point que, parfois, on ne sait plus de quoi l'on parle ! L'IA est un vaste machin ou tout est mis dans le même sac. On y anthropomorphise à outrance l'IA en imaginant qu'elle imite, remplace et dépasse l'homme. Cela donne parfois envie de remettre quelques pendules à l'heure !

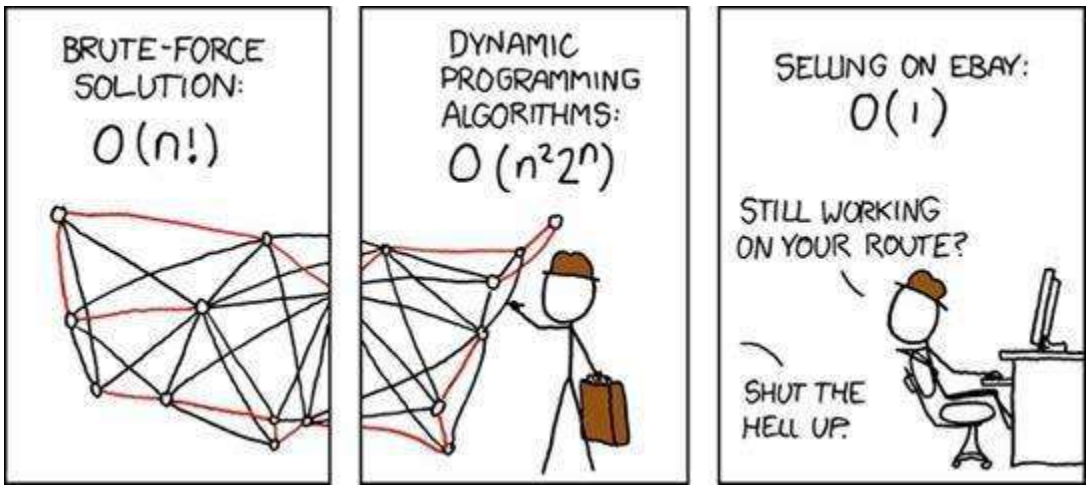
Sur ce, je vais maintenant partir des couches d'abstraction les plus basses (systèmes experts, réseaux neuronaux, machine learning, méthodes statistiques, …) pour ensuite monter dans les couches plus hautes qui font généralement appel aux couches basses, comme dans la reconnaissance de la parole ou des images. Pour chacune de ces techniques, je vais évoquer si besoin est leur ancienneté, les progrès les plus récents, les applications phares ainsi que quelques acteurs des marchés correspondants.

Force brute et algorithmes traditionnels

La force brute 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 peuvent être optimisés avec des algorithmes d'élagage qui évacuent les "branches mortes" de la combinatoire ne pouvant aboutir à aucune solution. C'est plus facile à réaliser aux échecs qu'au jeu de Go !

La force brute a été utilisée pour gagner aux premiers avec l'ordinateur **Deeper Blue** d'IBM en 1997, calculant 200 millions de positions par seconde. Des réseaux neuronaux ont été exploités pour gagner au Go récemment avec la solution créée par **DeepMind**, la filiale en IA de Google. Avec un mélange de force brute et de machine learning permettant de faire des économies de combinatoires à tester.

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 réseaux neuronaux ou de machine learning que nous traiterons plus loin. Elle ne fonctionne que si la combinatoire 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.



(**source de l'image**)

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.

Moteurs de règles et systèmes experts

Les débuts des moteurs de règles remontent à 1957 quand **Alan Newell** et **Herbert Simon** développaient le General Problem Solver (GPS), un logiciel de résolution de problèmes 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.

Les moteurs de règles s'appuient sur la notion de raisonnement contraint par des règles. On fournit au moteur un ensemble de règles 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*".

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.

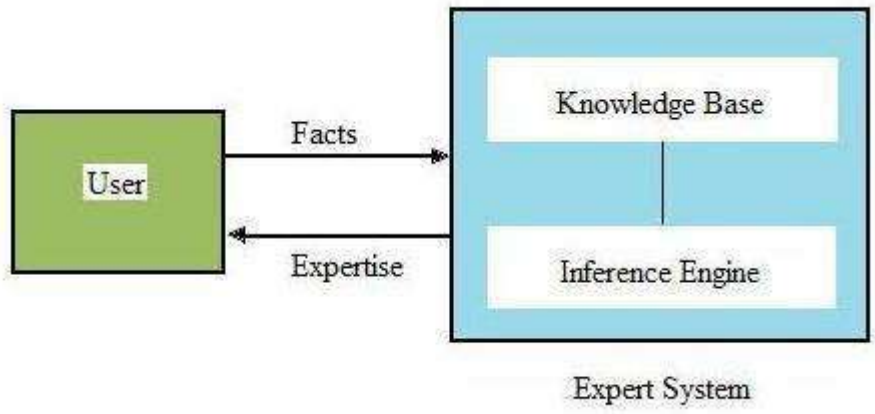
Cette branche de l'IA a été introduite par **John McCarthy** en 1958. Elle aboutit dans les années 1970 aux travaux de **Robert Kowalski** de l'Université d'Edinbourg, d'**Alain Colmerauer** et **Philippe Roussel** qui sont à l'origine du langage de programmation **Prolog** qui connut ses heures de gloire dans les années 1980.

Le langage **LISP** a été aussi utilisé dans ce domaine. Il s'est même développé une petite industrie avec les ordinateurs spécialisés de **Lisp Machines** et **Symbolics** (1979-2005), et des logiciels d'**Intellicorp** (créé en 1980 et maintenant spécialisé dans les logiciels de gestion d'applications pour SAP, un métier plus terre à terre).

Les moteurs de règles sont employés dans les systèmes experts, un domaine et un marché qui se sont développés depuis les années 1980. Les systèmes experts ont été notamment 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.

L'approche se heurtait cependant à la difficulté de capter la connaissance des experts. Cela explique son déclin dans les années 1990. Dans de nombreux domaines, la force brute s'est imposée en lieu et place de la logique et de la captation manuelle de connaissances.

Cela se retrouve dans le traitement du langage, la traduction automatique, la reconnaissance des images ou les moteurs de recherche. Même IBM Watson utilise la force brute pour exploiter de gros volumes de bases de données de connaissances non structurées.



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, puis un moteur d'inférence, 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.

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. 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 dans le prochain article de cette série.

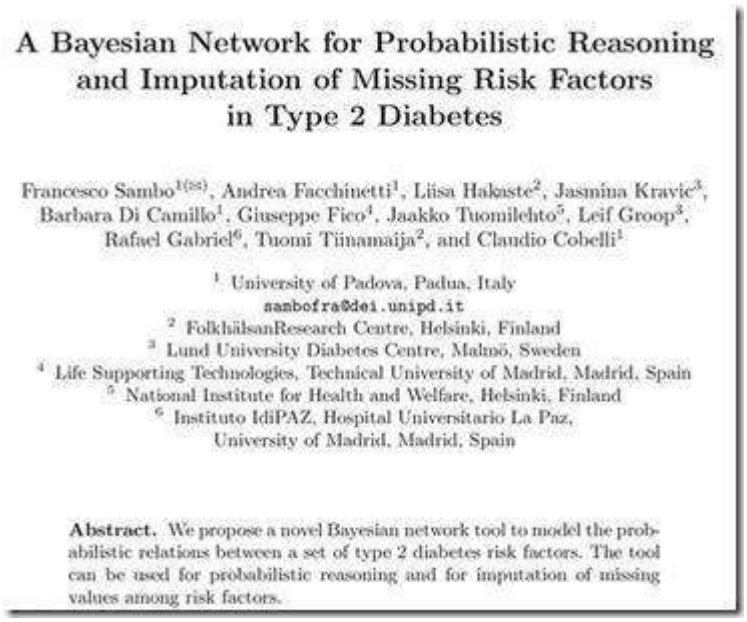
Méthodes statistiques

Les méthodes statistiques et notamment bayésiennes permettent de prévoir la probabilité d'événement 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.

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[3].



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

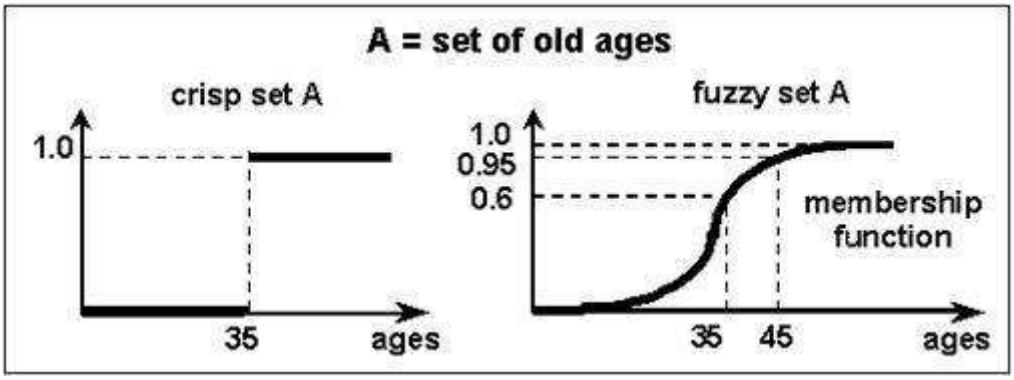
Abstract. We propose a novel Bayesian network tool to model the probabilistic relations between a set of type 2 diabetes risk factors. The tool can be used for probabilistic reasoning and for imputation of missing values among risk factors.

Logique floue

La logique floue est un concept de logique inventé par l'américain **Lofti Zadeh** ("Fuzzy Logic") en 1965. 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 !

La **logique floue** 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".

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



Quid des applications ? Elles sont relativement rares. 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.

Réseaux de neurones

Les réseaux de neurones visent à reproduire approximativement par bio mimétisme le fonctionnement des neurones vivants 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. Combinées en grand nombre, les neurones artificiels permettent de créer des systèmes capables par exemple de reconnaître des formes. Les réseaux neuronaux les plus intéressants sont ceux qui peuvent faire de l'auto-apprentissage. Attention cependant, les réseaux neuronaux visent l'efficacité algorithmique et ne prétendent pas être des reproductions fines du système nerveux biologique à bas niveau.

Le concept de base est né en 1943 des travaux de **Warren McCullochs** et **Walter Pitts**. **Donald Hebb** ajouta le principe de modulation la connexion entre neurones en 1949, permettant aux neurones de mémoriser de l'expérience. La connaissance est acquise via les interconnexions entre neurones et via un processus d'apprentissage. Elle est matérialisée sous la forme de poids de connexions synaptiques entre neurones qui varient en fonction de l'expérience acquise, par exemple dans la reconnaissance d'images.

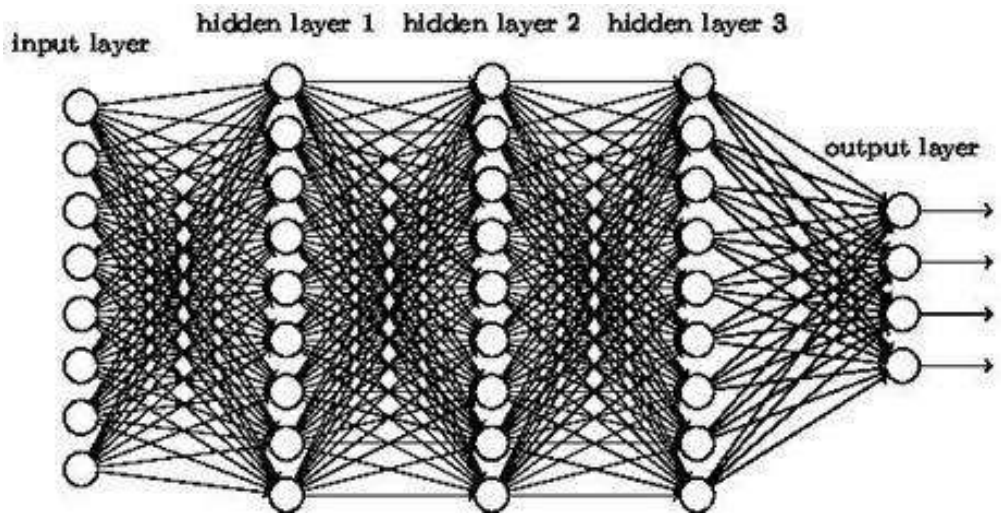
Le premier réseau de neurones matériel fut créé par **Marvin Minsky** et **Dean Edmons** en 1950. Le SNARC simulait 40 neurones avec 3000 lampes à tubes !

Frank Rosenblatt, un collègue de Marvin Minsky, créa ensuite le concept du **perceptron** en 1957 qui était un neurone assez simple dans son principe. 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 sévèrement les travaux de Frank Rosenblatt. D'ailleurs, sur un point très spécifique portant sur les portes logiques XOR des perceptrons. Un point valie et qui mit un coup d'arrêt à ces développements, un peu comme le rapport de Lightfill quelques années plus tard. Toujours, dans la dynamique de la rivalité des *neats* vs *scuffies*.

Ce coup d'arrêt 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 dans les années 1980, après le décès de Frank Rosenblatt.

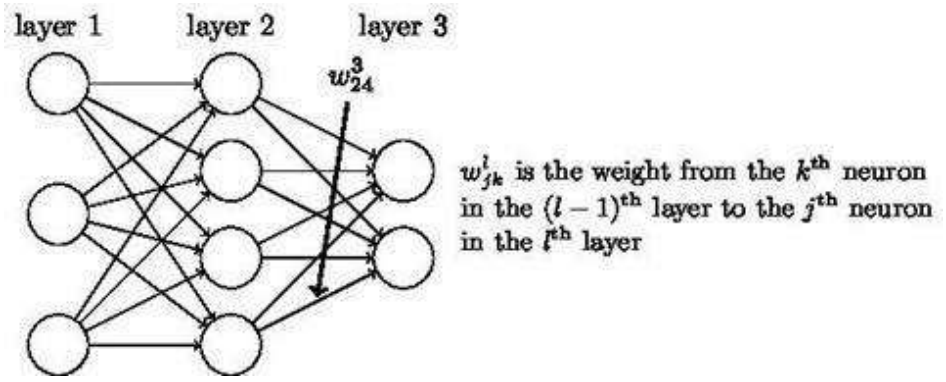
Depuis une vingtaine d'années, les réseaux neuronaux sont mis à toutes les sauces, la dernière étant la victoire de **DeepMind** contre le champion du monde de Go à la mi-mars 2016. Les réseaux neuronaux ont progressé pas à pas, avec la création d'innombrables variantes conceptuelles pour améliorer leurs capacités d'apprentissage et de mémorisation. 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. Les dernières années ont cependant vu les efforts de recherche passer des travaux dans la logique de base vers ses applications.



L'un des points clés des réseaux de neurones actuels est la technique de la **rétropropagation du gradient** (back propagation) qui corrige les défauts des Perceptrons identifiés par Papert et Minsky. Elle a vu le jour dans les **années 1960** puis, pendant et après le second hiver de l'IA, a repris son essor vers 1986.

Elle permet de modifier le poids des liaisons synaptiques entre neurones en fonction des erreurs constatées dans les évaluations précédentes, par exemple dans la reconnaissance d'images.

Comment fonctionne cette boucle d'apprentissage ? C'est un apprentissage soit assisté, soit automatique en comparant les résultats avec la bonne réponse, déjà connue. C'est un des débats clés d'aujourd'hui : est-on réellement capable de créer des réseaux doués de facultés d'auto-apprentissage ? Il semblerait que l'on en soit encore loin.



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.

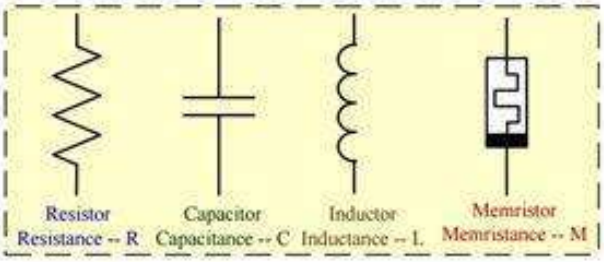
En 2011, son équipe développait un robot utilisant ces SOINN capable d'auto-apprentissage (**vidéo**), illustrant magistralement les applications des réseaux neuronaux. Nous sommes 10 ans plus tard, et on constate que les robots autonomes sont encore loin du compte, même si les sociétés telles que Boston Dynamics, filiale de Google, nous ébaubissent avec des robots très souples dans leur démarche et résistant à l'adversité.

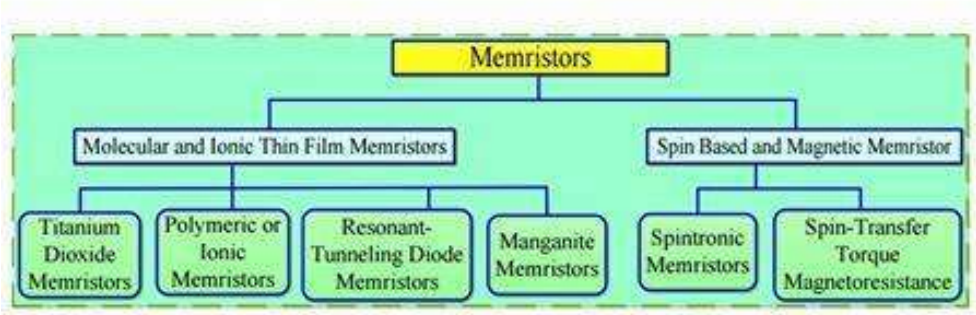
Les réseaux neuronaux ont aussi progressé grâce à leur mise en œuvre dans des architectures matérielles spécialisées permettant de bien paralléliser leurs traitements comme le fait le cerveau. Le composant électronique idéal pour créer un réseau de neurones est capable d'intégrer un très grand nombre de pico-unités de traitement avec entrées, sorties, logique de calcul si possible programmable et mémoire non volatile. Il faut par ailleurs que les connexions entre neurones (synapses) soient les plus nombreuses possibles. En pratique, les connexions se font avec les neurones adjacents dans les circuits.

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.

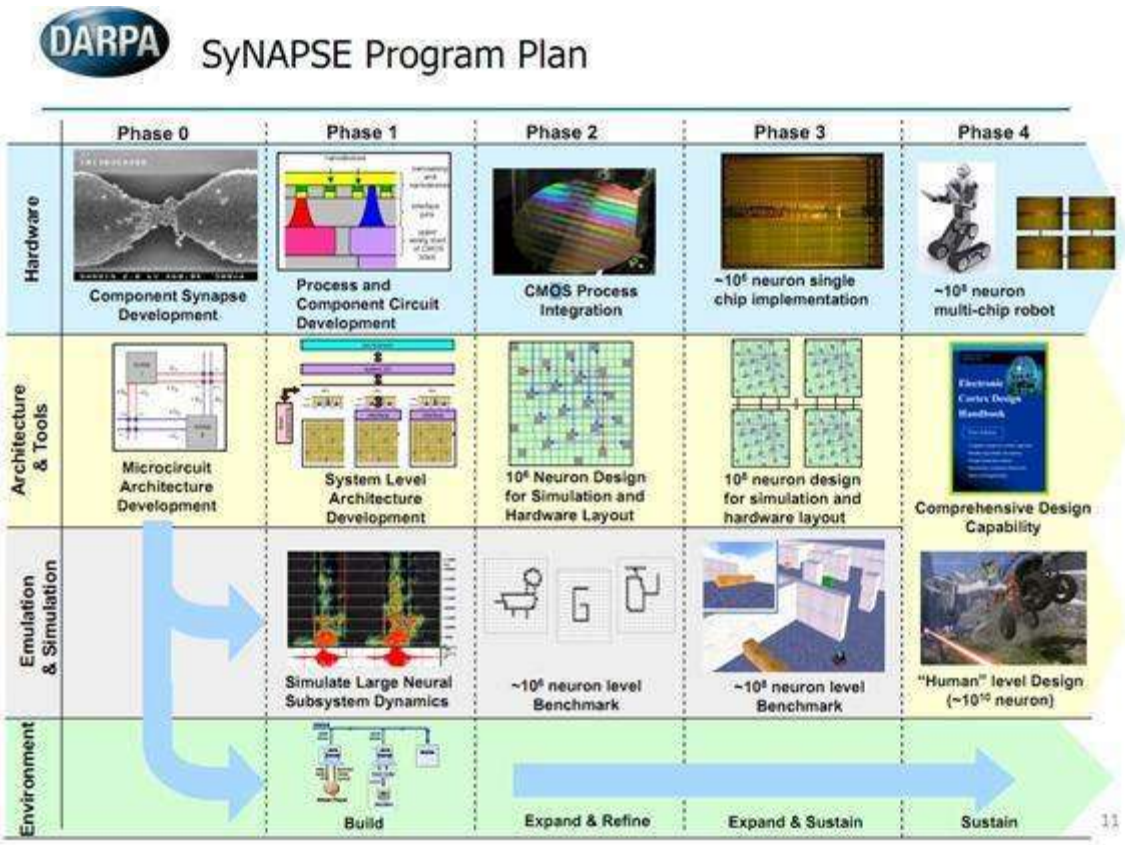
Ces composants sont intégrables dans des puces au silicium utilisant des procédés de fabrication plus ou moins traditionnels (**nanoimprint lithography**), en ajoutant une bonne douzaine d'étapes dans la production, et avec des matériaux rares comme les oxydes de titane.





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.
- 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

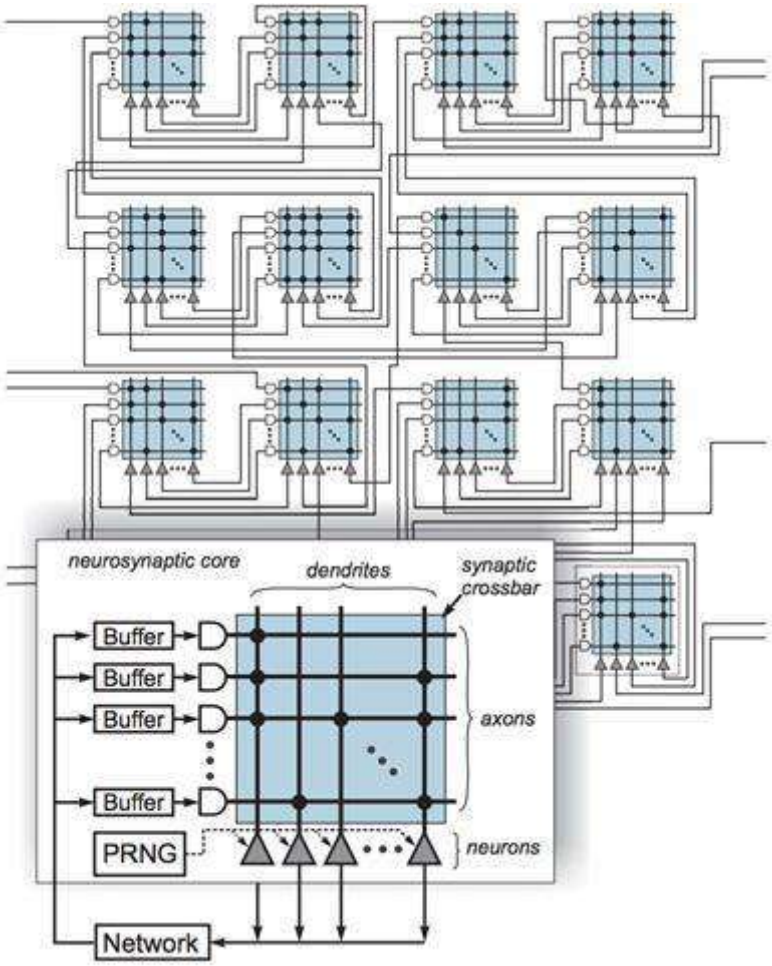
être terminé en 2013 et avait bénéficié d'une enveloppe de 740 K€. Difficile de savoir ce qu'il en est advenu en ligne.

- 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.

Le programme SyNAPSE de la DARPA a en tout cas 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.

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 (Skymake) réalisé en technologie 14 nm consomme entre 15 W et 130 W selon les modèles, pour 1,7 milliards de transistors.

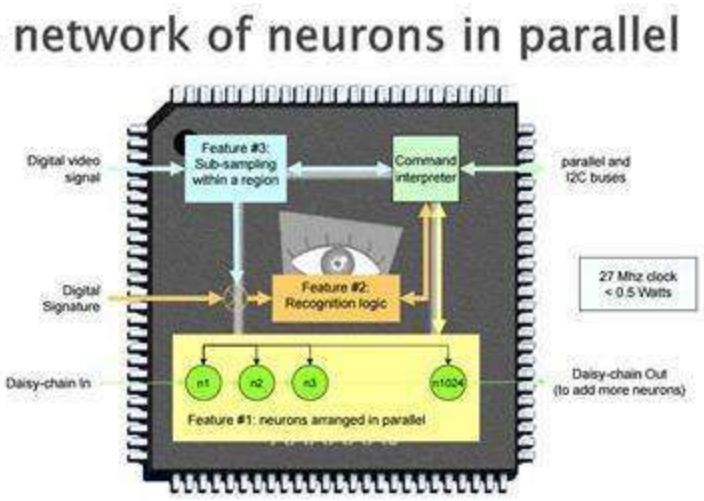
Le but d'IBM est de construire un ordinateur doté de 10 milliards de neurones et 100 trillions de synapses, consommant 1 KW et tenant dans un volume de deux litres. A titre de comparaison, un cerveau humain contient environ 85 milliards de neurones et ne consomme que 20 Watts ! Le biologique reste encore à ce stade une machine très efficace d'un point de vue énergétique !



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, celui de Stanford, qui travaille sur le chipset **Neurocore** intégrant pour l'instant 65536 neurones et fonctionnant à très basse consommation.

Il y a aussi le projet **SpiNNaker** de Steve Furber (Université de Manchester, UK), qui vise à créer un chipset de 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.

Enfin, dans le domaine commercial, le **CogniMem** CM1K est un chipset ASIC intégrant un réseau de 1024 neurones qui sert aux applications de reconnaissance des formes. Ne coutant que $94, il est notamment utilisé dans la **BrainCard**, issue d'une start-up française.



Plus récemment, **Nvidia** a 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. C'est bien mais probablement pas aussi optimal que de véritables réseaux de neurones et de synapses artificiels comme le TrueNorth d'IBM. Qui plus est, la carte PX2 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.

Il y aussi **Movidius** qui propose ses chipsets Myriad à base de réseaux neuronaux exploitant des processeurs vectoriels, dédiés au traitement de l'image. Ils en ont récemment lancé une version qui tient sur une clé USB, la Fathom Neural Compute Stick.

On peut donc constater que tout cela bouillonne, plutôt au niveau des laboratoires de recherche à ce stade, et que l'industrialisation prendra encore un peu de temps, mais que les réseaux neuronaux matériels ont probablement un bel avenir devant eux.

Support Vector Machines

Il faudrait aussi citer les **SVM** (Support Vector Machines) ou "Machine à vecteurs de support", une autre technique d'apprentissage supervisé plus adaptée à certaines

classes de problèmes que les réseaux neuronaux. Les SVM “scalent” mieux que ces derniers dans des architectures distribuées.

Ils peuvent servir par exemple à générer une classification automatique d’échantillons de données complexes, comme une segmentation client ou la classification de termes textuels. On les utilise aussi pour prédire la valeur numérique d’une variable. L’approche concurrence celle des réseaux bayésiens.

Machine learning et deep learning

Le vaste domaine du machine learning, ou apprentissage automatique, vise à faire des prédictions à partir de données existantes. C’est un domaine qui est intimement relié à celui des réseaux de neurones, qui servent de substrat pour les traitements. En effet, les outils de machine learning et de deep learning s’appuient sur différentes variantes de réseaux de neurones pour leur mise en œuvre pratique, notamment des réseaux neuronaux à plusieurs niveaux. Ces réseaux sont supervisés ou pas selon les cas.

Le machine 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 non structurées et à gérer des bases de connaissances. 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.

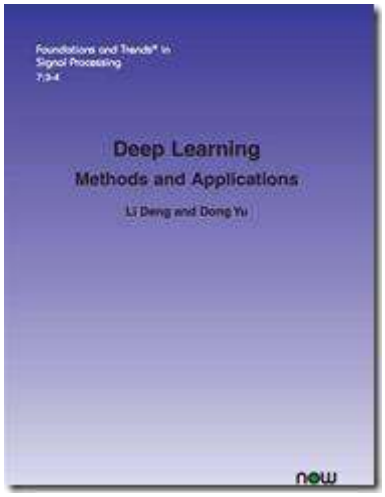
Les réseaux neuronaux ont connu un renouveau en 2006 avec les travaux des canadiens **Geoffrey Hinton** et **Simon Osindero** et du singapourien **Yee-Whye Teh** publiés dans **A Fast Learning Algorithm For Deep Belief Nets**, qui optimisent le fonctionnement des réseaux neuronaux multicouches.

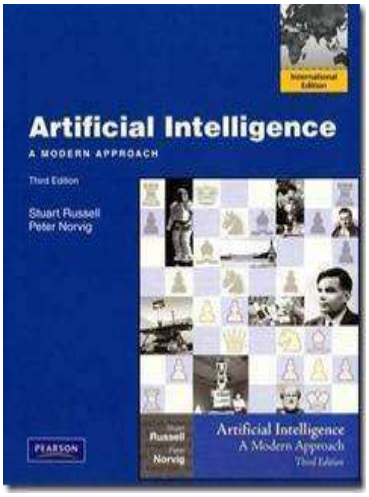
Le concept du machine learning a été ensuite formalisé par Geoffrey Hinton en 2007 dans **Learning multiple layers of representation**. Il s’appuyait lui-même sur les travaux du français **Yann LeCun** (en 1989) qui dirige maintenant le laboratoire de recherche en IA de Facebook et de l’allemand **Jürgen Schmidhuber** (1992) dont deux des anciens étudiants ont créé la start-up **DeepMind** maintenant filiale de Google. Petit monde !

Geoffrey Hinton travaille pour **Google** depuis 2013, pas loin du légendaire **Jeff Dean**, arrivé en 1999 et qui planche maintenant aussi sur le deep learning**.**

Pour comprendre le fonctionnement du deep learning, 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 !

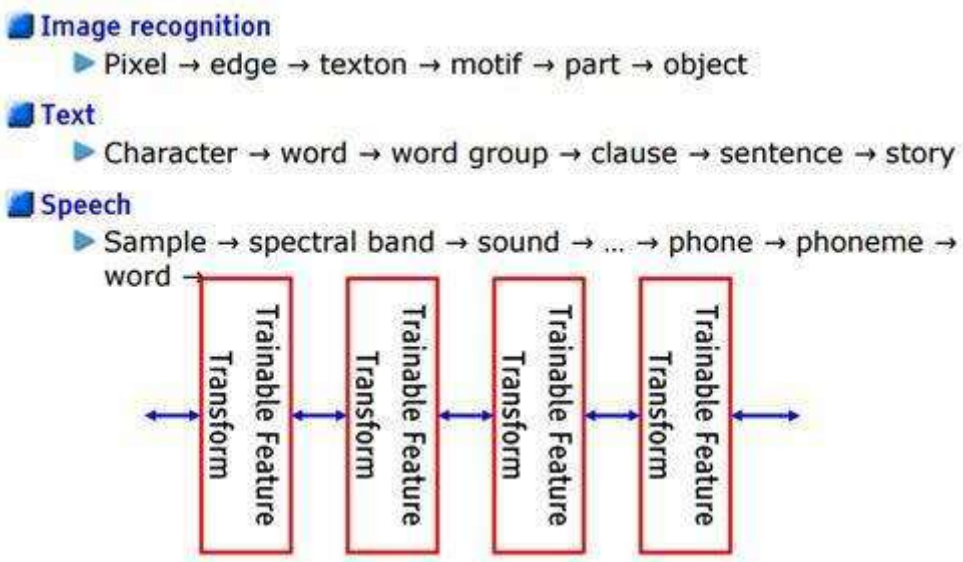
Il y a aussi **Deep Learning Methods and Applications** publié par Microsoft Research (197 pages) qui démarre en vulgarisant assez bien le sujet. Et puis **Artificial Intelligence A Modern Approach**, de Stuart Russel et Peter Norvig, une somme de référence sur l'IA qui fait la bagatelle de 1152 pages et qui ne serait que le B-A-BA pour les étudiants en informatique.





J'ai enfin trouvé cette présentation plutôt synthétique **A very brief overview of deep learning** de Maarten Grachten en 22 slides ! Ouf !

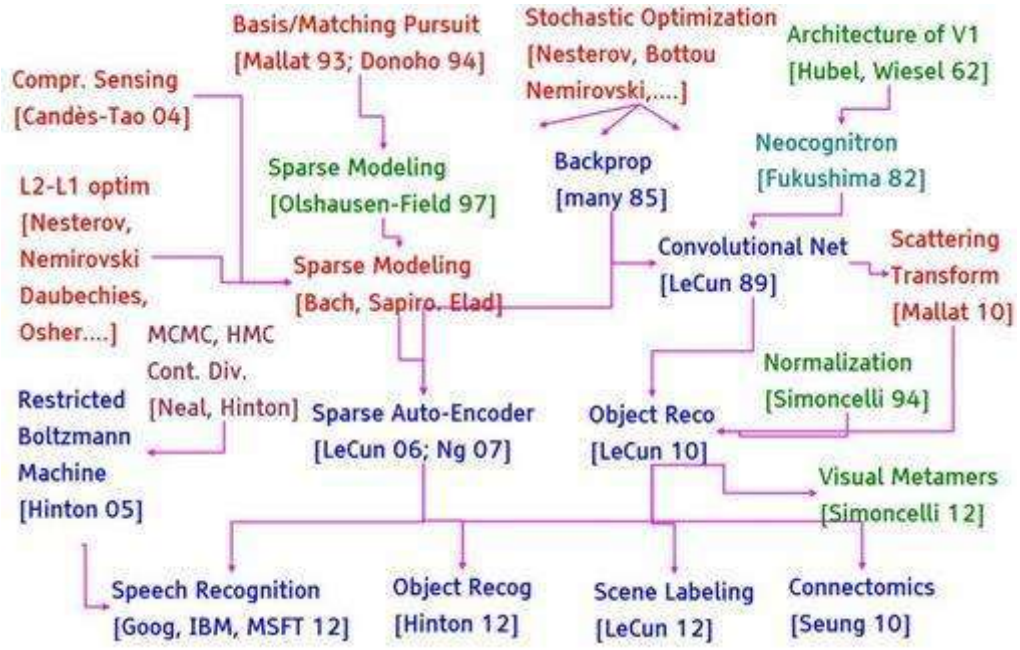
Bref, il faut se taper l'équivalent de plusieurs **Rapports du CES de Las Vegas** !



Le **deep learning** est une variante avancée du machine learning qui s'appuie sur des architectures en couches utilisant des "Restricted Boltzmann Machines" qui s'alimentent les unes les autres. Une machine de deep learning possède plus de couches dites cachées dans ses réseaux de neurones qu'une architecture de machine learning classique.

Le deep learning permet d'élever le niveau d'abstraction du machine learning en exploitant des concepts de plus haut niveau, comme indiqué dans le schéma ci-dessus [4]!

Voir aussi la **conférence inaugurale** de Yann LeCun au Collège de France en février 2016 où il excelle dans la vulgarisation). Ce fonctionnement imite d'ailleurs celui du cerveau humain qui utilise plusieurs niveaux d'abstraction. Il permet en théorie de mettre en place des méthodes d'auto-apprentissage sans base de données d'entrainement. Pour l'instant, cependant, cela semble surtout concerner les sens artificiels.



Cette complexité du sujet m'intrigue particulièrement. J'ai pu récemment explorer des sujets aussi variés que les réseaux M2M, la biologie moléculaire ou la génomique et ils m'ont semblé bien plus abordables que l'IA, même en parcourant des ouvrages de spécialistes ! C'est dire !

Cette difficulté d'appréhender la science derrière l'IA a probablement un lien avec les fantasmes qui lui sont associés. On imagine à fois le meilleur et le pire de ce que l'on ne peut pas expliquer avec son espace restreint de connaissances.

Cela permet aussi de mettre l'IA à toutes les sauces dans le marketing. L'appellation d'IA est encore utilisée pour valoriser certaines offres mais le machine learning l'est tout autant maintenant.

Derrière l'habillage marketing, il reste à comprendre ce que le fournisseur a réellement produit : a-t-il assemblé des briques logicielles existantes (souvent en open source), a-t-il créé des briques spécifiques, a-t-il juste entraîné un modèle, mise en forme des données, la solution est-elle une simple application directe de techniques existantes ?

[4] Source : l'excellent slideshow de **LeCun – Ranzato**, qui fait 204 slides, source également de l'historique du deep learning de cette page.

Reconnaissance de la parole

C'est une technologie commune et disponible dans les applications grand public. Le marché est dominé par de grands acteurs américains (OK **Google**, **Microsoft** Cortana, **Apple** Siri, **Amazon** Alexa, **Viv**[5]) et, en OEM, par l'américain **Nuance** qui vend sa solution un peu partout. 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.

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.

Elle profite aussi de l'augmentation régulière de la puissance des processeurs et notamment des processeurs mobiles. **Nvidia** propose depuis peu d'exploiter les fonctions GPU de ses chipsets pour mettre en œuvre des techniques de deep learning, bénéficiant du fort parallélisme des nombreux GPU disponibles.

Les solutions de reconnaissance vocale ont 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 se sent si on utilise un smartphone. 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.

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.

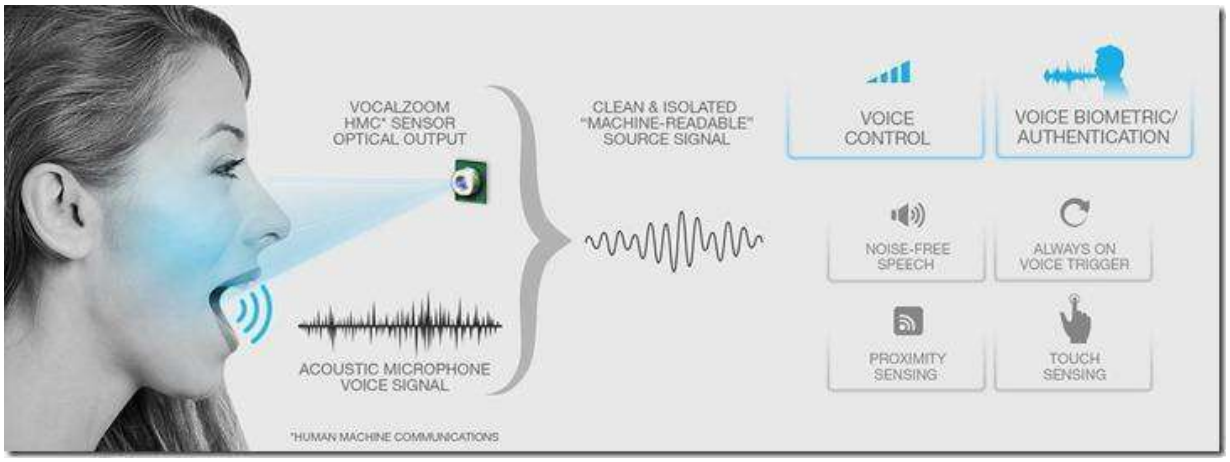
On est encore **loin de la solution parfaite**, notamment parce que les logiciels manquent d'informations sur le contexte des conversations[6]. Le taux de fiabilité n'est jamais de 100%. Il ne l'est d'ailleurs jamais pour l'homme également ! Par exemple, Microsoft Cortana atteint un taux d'erreur d'environ 8%, soit le double de celui de l'homme. Ce taux d'erreurs de l'IA diminuerait de 25% par an. Microsoft prévoit d'atteindre le taux d'erreur humain d'ici quelques années. Et encore, c'est pour l'anglais ! Le taux d'erreur 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**.

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 existent aussi. 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

[5] 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 de une seconde qui va la traiter. Viv a été présenté récemment lors de TechCrunch Disrupt à New York (vidéo).

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

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.



Reconnaissance d'images

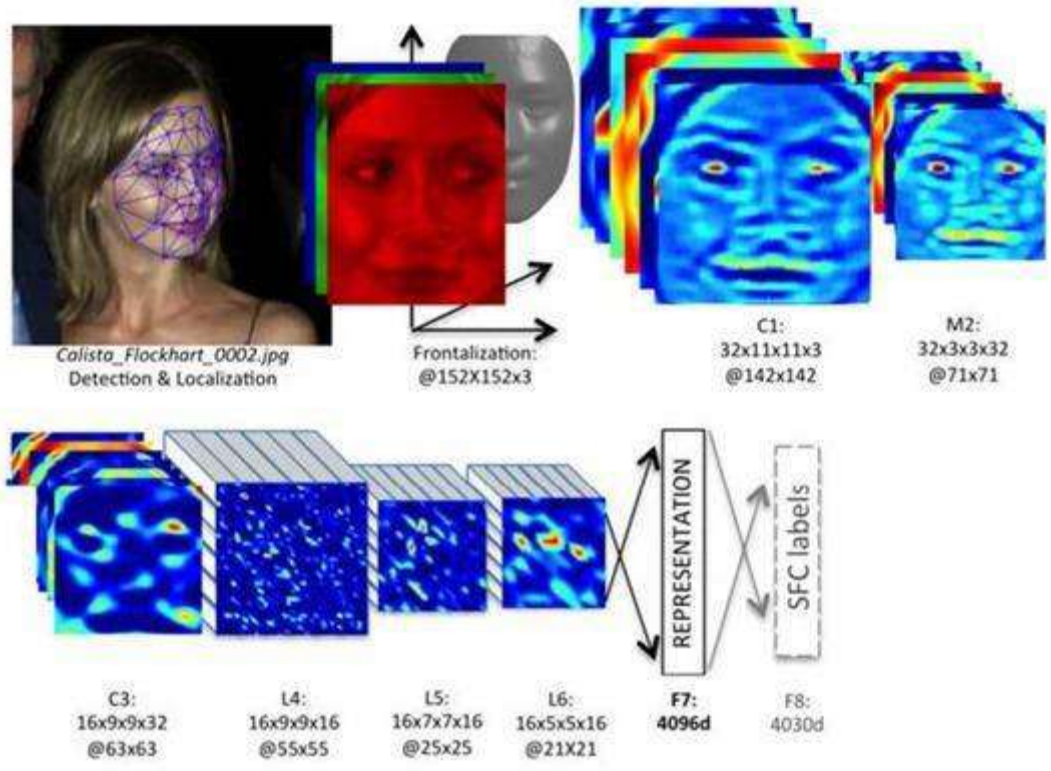
Cette fonction est devenue aussi critique que la reconnaissance de la parole et notamment dans les moteurs de recherche et certains réseaux sociaux, pour identifier des visages, des expressions ainsi que des lieux. Elle est aussi présente depuis des décennies dans les logiciels d'OCR pour reconnaitre les textes, 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.

La reconnaissance d'images est l'une des principales applications des réseaux neuronaux, du machine learning et du deep learning comme nous l'avons vu précédemment. 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. Par exemple, dans le cas des solutions de **Nvidia** ou **Mobileye** pour la conduite assistée, il s'agit de détecter au pixel près ceux qui correspondent à des piétons, des cyclistes, des véhicules, de la signalisation au sol et des panneaux de signalisation.

Google est évidemment friand de ce genre de technologies qu'ils utilisent dans Google Image et Google Photo. Google Image est capable (avec le glisser-déplacer) d'identifier des images similaires à celle que l'on fournit. Cela utilise probablement une méthode simple de création de hash-code sur les photos et de recherche dans l'index d'une grande base de données.

Dans leur projet FaceNet, Google annonce avoir atteint un taux de réussite de détection de visage de 99,63%[7]. Le tout s'appuie sur un réseau neuronal à 22 couches.

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



De son côté, **Facebook** et son projet DeepFace s'appuie sur la technologie issue d'une start-up israélienne **face.com**. Son taux de réussite serait 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%.

On trouve de la détection de visages dans plein de solutions du marché comme avec la fonction Faces de **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. 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** (seulement $300K de levés) et ses API en cloud de tagging automatique d'images en fonction de leur contenu, **Cognitec** qui vise surtout les marchés de la sécurité, **Cortexica** ($6,6m de levés) et son logiciel findSimilar en cloud qui met en œuvre ces techniques pour le retail et **Slyce** qui cible aussi le marché du retail ($37m de levés et IPO en avril 2015).

Citons enfin un domaine connexe, celui de 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**.

Reconnaissance de vidéos

La reconnaissance de vidéos est une évolution naturelle de la reconnaissance d'images, à ceci près que les vidéos fournissent plus d'information. Elle elle est utile dans tout un tas de contexte, notamment pour les voitures à conduite automatique, un domaine où l'ordinateur **peut maintenant dépasser l'homme**.

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 !

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 **Clarifai** ($10m de levés) qui permet notamment de faire de la curation de contenus photo et vidéo, ou chez le japonais **NEC**. Il faut aussi citer **OpenCV**, une solution open source de détection de visages. Voir **cette liste** de solutions pour développeurs de détection de visages dans les vidéos.

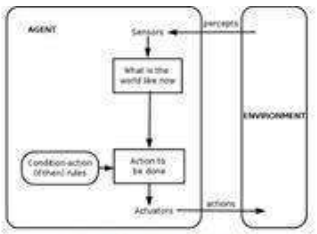
Agents intelligents et réseaux d'agents

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

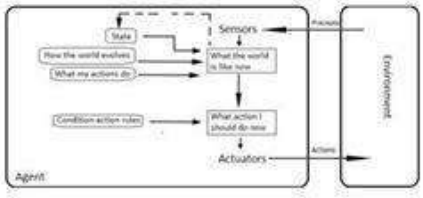
On compte notamment les **Distributed Problem Solving** (DPS) qui découpent un problème en sous-problèmes qui sont résolus de manière coopérative entre plusieurs agents reliés les uns aux autres. Ces systèmes sont conçus pour résoudre des problèmes bien spécifiques.

Les 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 :

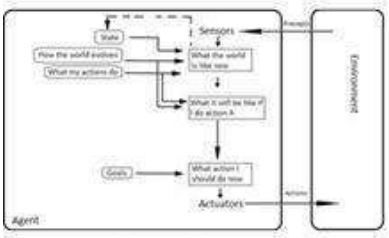
- Les **simple reflex agents** qui comprennent des capteurs, des règles indiquant quelle action mener et des actuateurs pour les déclencher. Ils travaillent en temps réel.
- Les **model based reflex agents** qui ajoutent un moteur d'état capable de mémoriser dans quel état se trouve l'objet et qui évaluent 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.



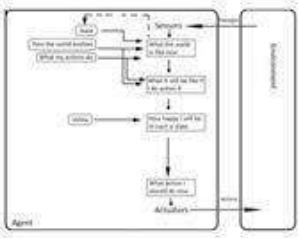
Simple reflex agent



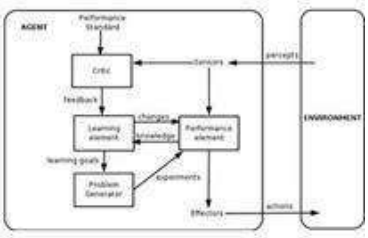
Model based reflex agent



Model-based, goal-based agent



Model-based, utility-based agent



General learning agent

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.

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 traduc-

tion 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.

Le robot Nao d'Aldebaran a une belle capacité de mouvement grâce à une mécanique de bon niveau. Il interagit en parlant avec l'utilisateur, mais de manière encore limitée. Son grand frère Pepper, est doté d'une capacité à capter les émotions des humains qu'il a en face de lui mais sa capacité de dialogue est encore approximative.

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, 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.

Le niveau d'abstraction des réseaux d'agents est plus élevé que celui des réseaux de neurones. D'où le fait que j'en termine par là sur cette partie !

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 ! Quid des outils de développement associés ? **Il y en a plein**, et notamment en open source.

IBM Watson et le marketing de l'intelligence artificielle

Je vais m'attaquer ici au marketing des offres de l'intelligence artificielle. Nous allons commencer par étudier le cas d'IBM avec Watson puis élargir la réflexion au marché en général, et notamment des startups qui surfent sur la vague de l'IA.

La prouesse technique et marketing d'IBM Watson

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 peur remontée par leurs clients qui avaient peur de perdre leur poste de management. Et oui ! Comme aujourd'hui, où l'on entend les prospectivistes annoncer l'Armageddon dans plein de catégories d'emplois. Comme nous l'avons vu dans l'article précédent, ces craintes étaient alimentées par des prévisions un peu trop optimistes côté timing issues d'experts de l'IA comme Herbert Simon. Prenons-en de la graine [8]!

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, au moment de l'arrivée de son nouveau CEO de l'époque, Lou Gerstner. Aujourd'hui, IBM est une société à nouveau en déclin, en tout cas en termes de chiffres d'affaires. Celui-ci est passé sous celui de Microsoft en 2015, une belle barre symbolique, surtout dans la mesure où Microsoft n'est même plus l'étalon de la croissance dans le numérique depuis que Google, Facebook et Apple lui ont damé le pion.

En tout cas, 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 du Oracle, du Microsoft, du SAP, bref de tout, en fonction des contraintes du client.

La question reste cependant pour tout acteur du marché de ne pas rater les vagues technologiques. IBM s'en était pas trop mal sorti en 2000 en se positionnant dans le e-business. Sa campagne de communication martelait le rôle de "one-stop-shop" provider d'IBM pour ses clients. Pour la petite histoire, elle avait été pilotée par un certain Pierre Chappaz, devenu ensuite créateur de Kelkoo, puis Wikio, intégré depuis dans le groupe Teads.

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 architec-

[8] Source : **Humans Need not Apply** – 2015, de Jerry Kaplan.

tures, 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 bataille des jeux d'échecs. Durant plusieurs années, elle culmina avec la victoire de l'ordinateur IBM Deeper Blue (*ci-dessus*[9]) 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. Trois parties avaient été organisées : un échauffement le 13 février, et deux les 14 et 15 février. J'ai regardé l'épisode du 14 février **sur YouTube**. Il mettait face à face un avatar représentant Watson sur écran et deux champions du jeu : Brad Rutter et Ken Jennings. Dans les trois parties, Jeopardy gagne, fine. On se rend compte que Watson ne répond pas bien à toutes les questions, tout du moins au début. C'est un diésel ! Il est lent au démarrage mais carbure bien ensuite.

J'ai aussi trouvé une autre **partie intéressante** et moins médiatisée organisée avec comme joueurs Miles O'Brien et David Gondek, l'un des créateurs de Watson. Au départ, Watson ne sait pas indiquer pendant quelle décennie Klaus Barbie a été con-

[9] Deep Blue ou Deeper Blue ? Deep Blue a en fait connu plusieurs versions et celle qui a gagné était dénommé Deeper Blue.

damné. OK, ce n'est probablement pas un problème de compréhension mais plutôt de bases de connaissances utilisées ! Il ne sait pas indiquer sur quelle place de Dallas (Dealey Plaza) JFK a été assassiné. La littérature sur le sujet est pourtant abondante aux USA. Il ne sait pas non plus ce qu'est la vermiphobia (la phobie des vers) ni la ailuraphobia (phobie des chats) dont la signification est disponible sur Wikipedia. Mais la reconnaissance vocale est peut-être défaillante sur ces termes peu usités.

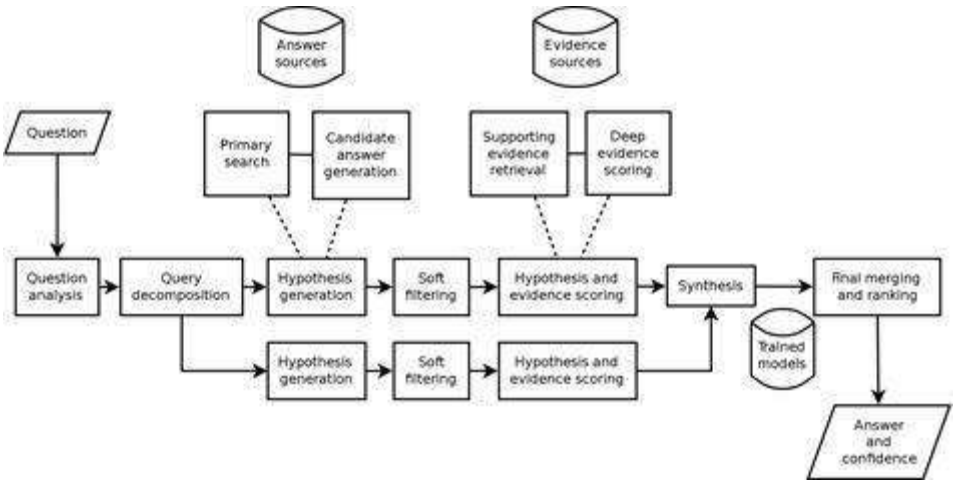
Il ne sait pas non plus identifier des recettes de cuisine en fonction de leurs composantes. C'est plutôt décevant. Un peu comme les systèmes de reconnaissance de la parole qui marchotent en situation difficile. Watson a du mal à répondre à des questions formulées avec peu de mots et comprenant des ambiguïtés ou des doubles sens. Watson ne devait pas non plus accéder à des sources d'information suffisamment larges, et en particulier étrangères. Pourtant, il s'agissait tout de même de 200 millions de pages de données structurées et non structurées représentant un total de 4 To. Mais j'ai lu quelque part que dans Jeopardy, les questions n'étaient pas interprétées par Watson par reconnaissance de la parole mais pas saisie humaine. Ce qui induisait potentiellement une source d'erreur externe à Watson.

Pendant la première moitié de la partie, Watson est en retard par rapport à ses deux concurrents. Il commence à battre ses concurrents à partir d'une série de questions visant à déterminer quelle capitale est la plus au nord parmi deux villes. Comme les candidats américains ont l'air d'être nuls en géographie (moi, m'sieu, je savais…), ils ne répondent à aucune de ces questions alors que Watson a une belle base de donnée en mémoire avec les villes et leurs coordonnées géographiques. Dans les questions qui suivent, Jeopardy prend le dessus. On se demande si l'impact psychologique de la remontée de Jeopardy avec les questions sur la géographie joue un rôle. Cette partie méconnue dure moins de 15 minutes en tout. L'un des deux joueurs a empoché $28000 et Watson, $31999. Ils sont au coude à coude.

Il a fallu du temps pour aboutir à ces performances de Watson ! Plusieurs parties d'essais avaient été perdues par Watson avant l'épisode historique de 2011. IBM indique que Watson s'est amélioré au gré des parties, mais il est difficile de faire la part des choses entre les évolutions logicielles, un éventuel auto-apprentissage et la fourniture de nouvelles sources de données.

En tout cas, le champion de Jeopardy Ken Jennings témoignait en février 2013 **dans TEDxSeattle** de cette douloureuse impression de devenir obsolète. A vrai dire, il ne l'est **toujours pas**.

Cinq ans après, le jeu Jeopardy est toujours diffusé sur la TV US et depuis 1964, un record ! De quoi relativiser les prévisions alarmistes sur l'obsolescence des métiers ! Même si joueur de Jeopardy n'est pas à proprement parler un métier.

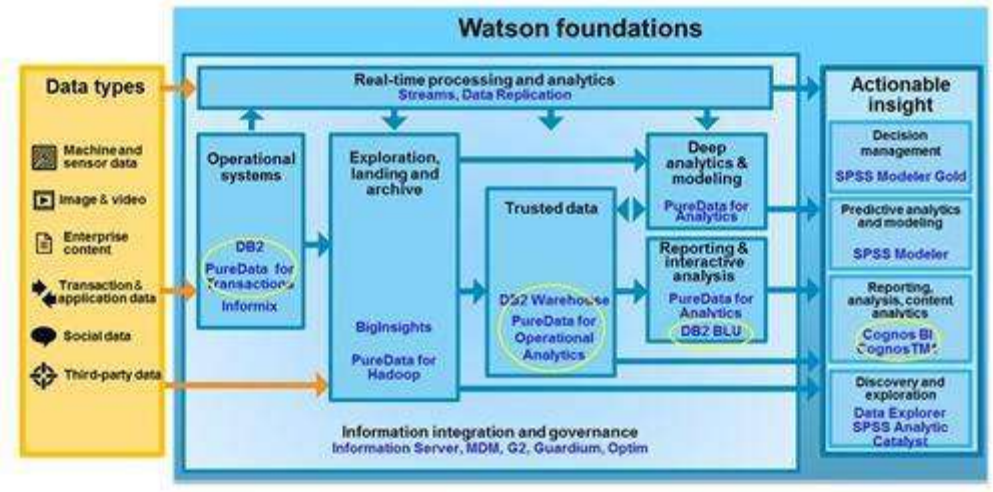


Cette histoire de Jeopardy est un peu enjolivée et construite par la communication d'IBM qui au passage, a été pilotée à l'échelle mondiale par l'agence **Ogilvy**. La performance n'est pas si impressionnante que cela. J'aurais été bluffé si Watson avait essayé de répondre à **toutes** les questions et écrasé ses concurrents à plates-coutures.

Cela montre la difficulté d'accéder à des informations faiblement structurées, même si tout est disponible sur Internet. Watson chargeait toute sa base de données interne en RAM et ne tapait pas en temps réel dans Internet ou dans un moteur de recherche pour des raisons de rapidité, d'où ses limites. Qui plus est, il semble que les questions étaient saisies au clavier par des opérateurs, et non pas traitées par reconnaissance vocale.

A priori cependant, les limitations rencontrées pendant la partie de 2011 ont pu être levées depuis. Ce n'est qu'une question de moyens. Il faudrait probablement organiser une partie "retour" où Watson répondrait (juste) à 100% des questions ! Et trouver des figurants pour le contrer ! Watson fait ce que Google Search devrait peut-être faire un jour, en allant plus loin que ses habituelles méthodes statistiques. Seulement voilà, Watson n'est pas aussi scalable que l'architecture de Google Search ! Pour l'instant. Sinon, Google mettrait probablement en œuvre des méthodes voisines de celles de Watson dans son moteur de recherche. Il en maîtrise très bien les composantes technologiques.

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, ou en dialoguant avec une machine, on ne sait pas distinguer l'homme de la machine. Watson était d'abord présenté comme un ordinateur. Il s'appuie 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. Cette fameuse RAM chargeant tout le corpus de contenus utilisé pour Jeopardy.



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**. 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.

Notons que la communication d'IBM est en tout cas très focalisée sur le cognitive computing et que Watson est le produit d'appel phare. C'était le seul sujet 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[10].

L'approche écosystème de Watson

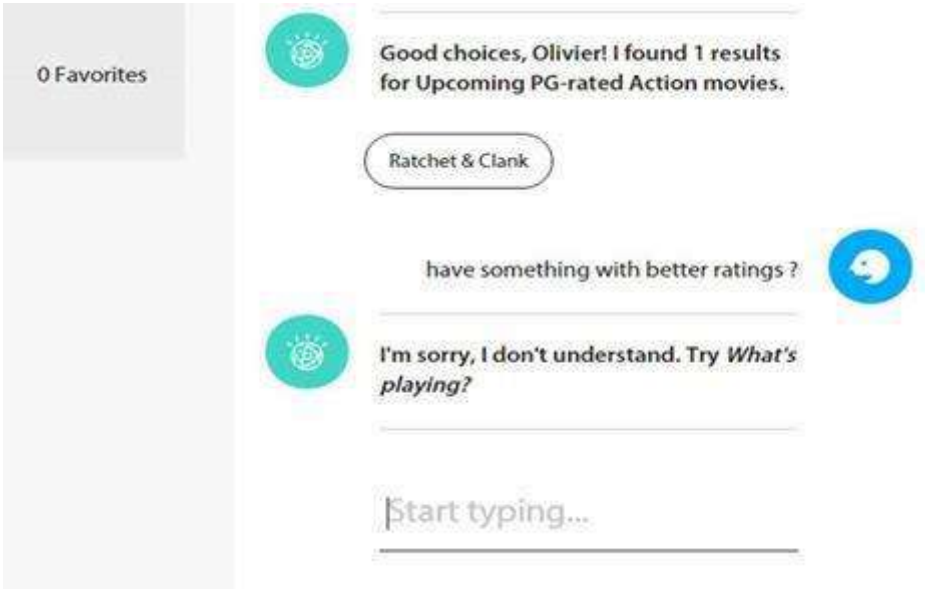
A chaque solution d'IA, son assemblage de composants hétéroclites réalisé sur mesure pour répondre à un besoin. C'est particulièrement vrai d'IBM Watson. Ce dernier est un très bon coup business et marketing d'IBM, qui a réussi à simplifier un sujet très complexe. Ils ont ainsi vulgarisé les capacités de Watson et pu cacher sa complexité, voisine de celle de l'architecture de WebSphere et dans laquelle cohabitent une belle part des techniques évoquées dans la partie précédente.

[10] Les vidéos des keynotes de l'édition de mai 2015 sont disponibles pour la **première** et la seconde journée.

IBM a 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 !

Fredrik Stenbeck a bien **décrypté** début mars 2016 ce que contient IBM Watson. Il est proposé aux développeurs de solutions sous la forme d'APIs REST[11] qui permettent d'accéder à la panoplie de services suivants :

- **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.
- Le **Natural Language Classifier** qui permet de classifier automatiquement des données textuelles, issues en général de questions posées par des clients en langage naturel.
- **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. Voici un exemple de **code** et le résultat associé :



- **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.

Créer une application Watson revient donc à créer du code, du contenu et à réaliser un travail d'intégration pour créer un agent conversationnel intelligent. Dans des ap-

[11] Avec requêtes http comprenant des GET et des POST et renvoyant le résultat.

proches verticales, il faut définir des scénarios de dialogues assez précis et avoir sous la main beaucoup de données exploitables, aussi bien structurées que non structurées.

D'où l'importance pour IBM d'avoir un écosystème de partenaires solutions à même de couvrir les besoins de divers marchés verticaux. Pour ce faire, IBM a lancé un programme partenaire assez classique qui comprend l'accès aux APIs, à une communauté, un programme d'accélération de trois mois et un catalogue de solutions pour promouvoir les partenaires. A ce jour, l'écosystème d'IBM Watson comprend environ 400 sociétés. 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 comprend maintenant 2000 personnes ans le monde, principalement des consultants, avant-vente et développeurs. Y compris, un centre d'avant-vente et de support situé à Montpellier.

L'ensemble est intégré dans les "IBM Cognitive Business Solutions" avec un focus sur quelques marchés clés : l'assurance, le retail et la santé. Ces 2000 personnes sont un bon début mais encore peu au regard des 270 000 collaborateurs d'IBM Services (c'est une estimation au doigt mouillé). 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.

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. L'impression est donnée d'un spectre fonctionnel assez limité et focalisé sur la création d'agents conversationnels. Les offres ne sont pas applicables au traitement de l'image ni à la robotique. C'est accentué par le fait qu'IBM ne communique pas dans le détail 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).

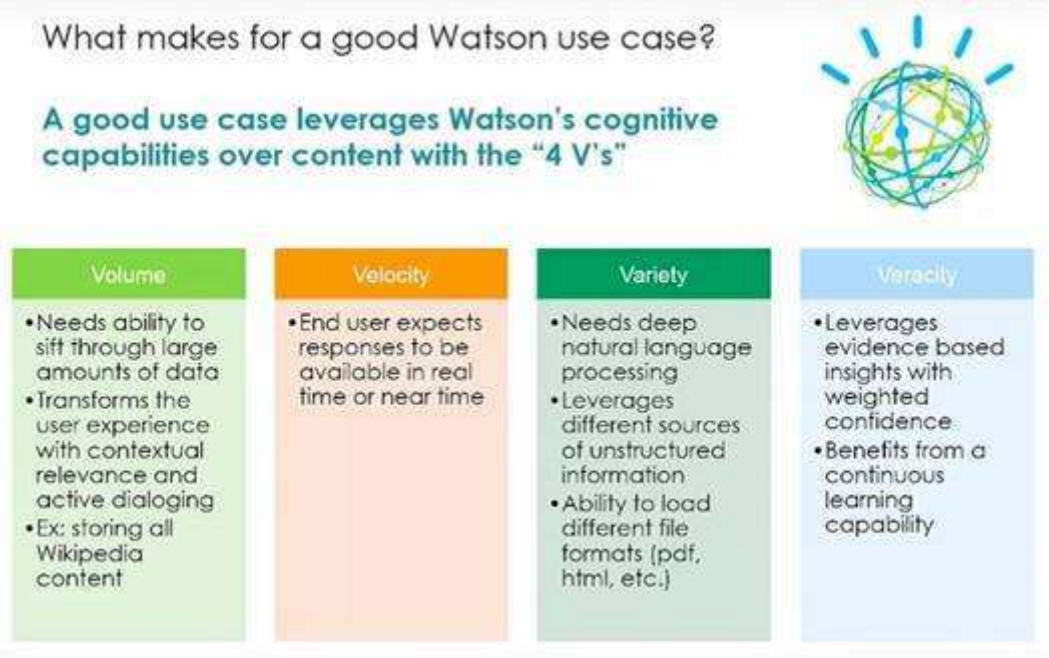
Avec leur approche service et intégration, IBM pourra cependant toujours affirmer qu'il sait intégrer les autres briques du marché. Bref, Watson est, en l'état, un objet difficile à benchmarker avec sa concurrence !

Les études de cas et projets d'IBM Watson

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 es 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.



Quand on observe le marketing et les études de cas avancées par IBM, les mêmes ont tendance à revenir systématiquement, notamment le système expert d'aide au diagnostic et de prescription pour le traitement de cancers. Leur montée en puissance commerciale est difficile à évaluer.

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.

Voyons donc ce qu'IBM a dans sa besace de références clients et partenaires, par segment de marché.

Santé

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.

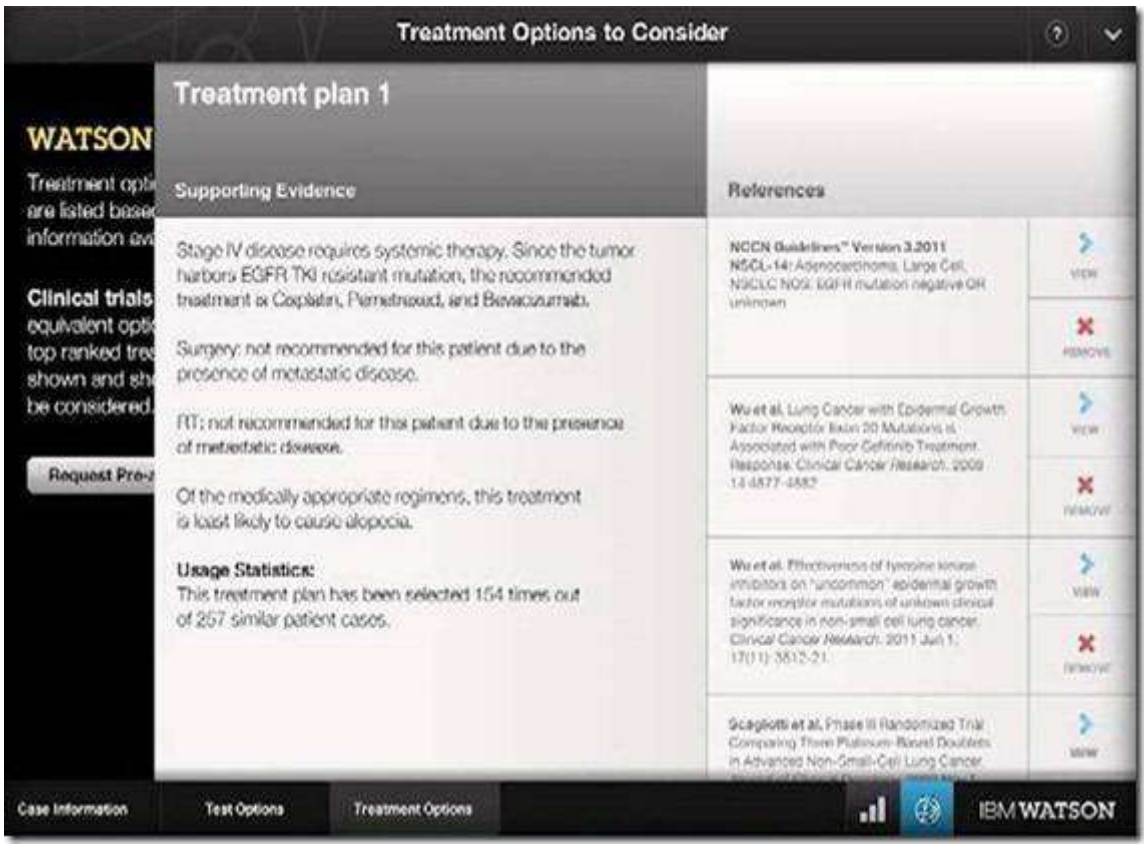
Elle a ensuite été déployée dans plus d'une quinzaine d'établissements aux USA et ailleurs dans le monde comme en Inde. Elle est fournie sous forme de service en cloud, avec un abonnement dont le prix n'a pas été rendu public par IBM.

La solution analyse les dossiers de patients atteints de tumeurs cancéreuses, y compris le séquençage d'ADN des tumeurs[12], aide au diagnostic, détermine des traitements possibles et évalue leur efficacité relative. Il aide notamment à optimiser l'usage de la chirurgie, de la radiothérapie et de la chimiothérapie. Les cancers sont des pathologies idéales pour Watson car elles sont plurifactorielles. Mais ce n'est pas (encore) de la médecine préventive.

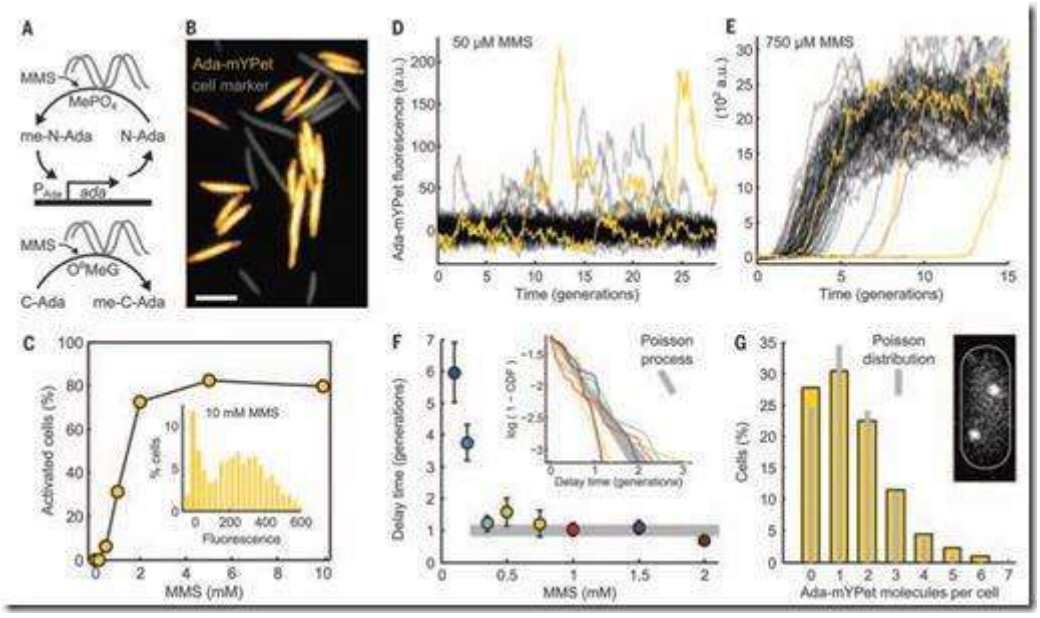
Les études scientifiques publiées sont très nombreuses et toujours fournies avec des résultats statistiques sur des cohortes de patients. Il faut les croiser avec des logiques statistiques bayésiennes et cognitives complexes pour en tirer des conclusions. On connait par exemple le lien entre les mutations 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 données relationnelles, etc). Il semble que cette partie de la solution ait été développée en partenariat avec **Cleveland Clinic**.

[12] 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.



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.

Watson fournit au praticien un choix de traitements qui sont fournis avec un indice de confiance, comme la probabilité de survie. Après avoir démarré avec les cancers du poumon, les cancers couverts intègrent maintenant les leucémies, les mélanomes, ceux du pancréas, des ovaires, du cerveau, du sein et du colon.

Dans cette application, Watson bat l'homme dans la force brute : il compulse notamment des bases de données de recherche en oncologie pour aider les cancérologues. Mais d'où viennent ces données ? Fait-il progresser la recherche ? Indirectement oui car il va alimenter ces bases de données qu'il utilise avec des résultats de traitement saisis par les praticiens.

Par contre, il ne fait pas directement progresser la recherche sur les cancers. Il ne faut pas oublier que chacun des articles scientifiques exploité a nécessité de 3 à 7 années de recherche 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.

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.

Distribution

Dans le retail, 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 **Fluid**. 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[13].

L'éditeur de logiciel américain **Red Ant** a aussi développé une solution de formation des commerciaux, SellSmart, qui accède au CRM de l'enseigne utilisatrice (**vidéo**).

Juridique

Le secteur juridique exploite de très gros volumes de données qualitatives : les lois et réglementations dans chaque pays, la jurisprudence des tribunaux et la littérature qui les commente. C'est un secteur qui fait appel comme la médecine à des praticiens qui doivent mémoriser de grandes quantités de textes. Watson arrive à point nommé pour les aider à piocher dans l'immensité du savoir de leur profession. Toujours pour les assister plus que pour les remplacer, sauf peut-être pour les tâches les plus élémentaires.

De nombreuses solutions juridiques sont déjà bâties sur Watson, surtout aux USA. Nous avions eu l'occasion d'en décrire certaines dans **Les disruptions numériques dans les professions libérales** en février 2016.

On compte notamment **LegalZoom,** un service d'avocat en ligne couvrant à la fois le droit des affaires et le droit civil[14] et l'avocat virtuel de la start-up **Ross Intelligence**, un agent conversationnel capable de répondre à un éventail varié de questions juridiques, adopté en mai 2016 par le cabinet d'avocats américain **Baker Hostetler**.

Le débat reste évidemment ouvert dans les professions juridiques pour évaluer la portée des solutions construites sur Watson et l'impact qu'il aura sur les métiers. Comme d'habitude, c'est la partie de ces métiers qui est la plus commoditisée et répétable qui sera automatisée en premier. Pour la partie qui relève plus de la dimension humaine, comme dans pénale, on aura encore besoin d'avocats pendant pas mal de temps. Et heureusement, si l'on parle d'automatiser une partie du travail des avocats, on ne l'évoque pas encore pour ce qui est des juges ! Un scénario avocat-IA contre juge-IA serait en tout cas intéressant à tester !

Divers

Watson a aussi fait son apparition sporadique dans divers marchés :

Dans les assurances, avec **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.

---

[13] Pour en savoir plus voir ce compte-rendu détaillé sur le JDN : **Comment The North Face a appliqué Watson à l'expérience d'achat**.

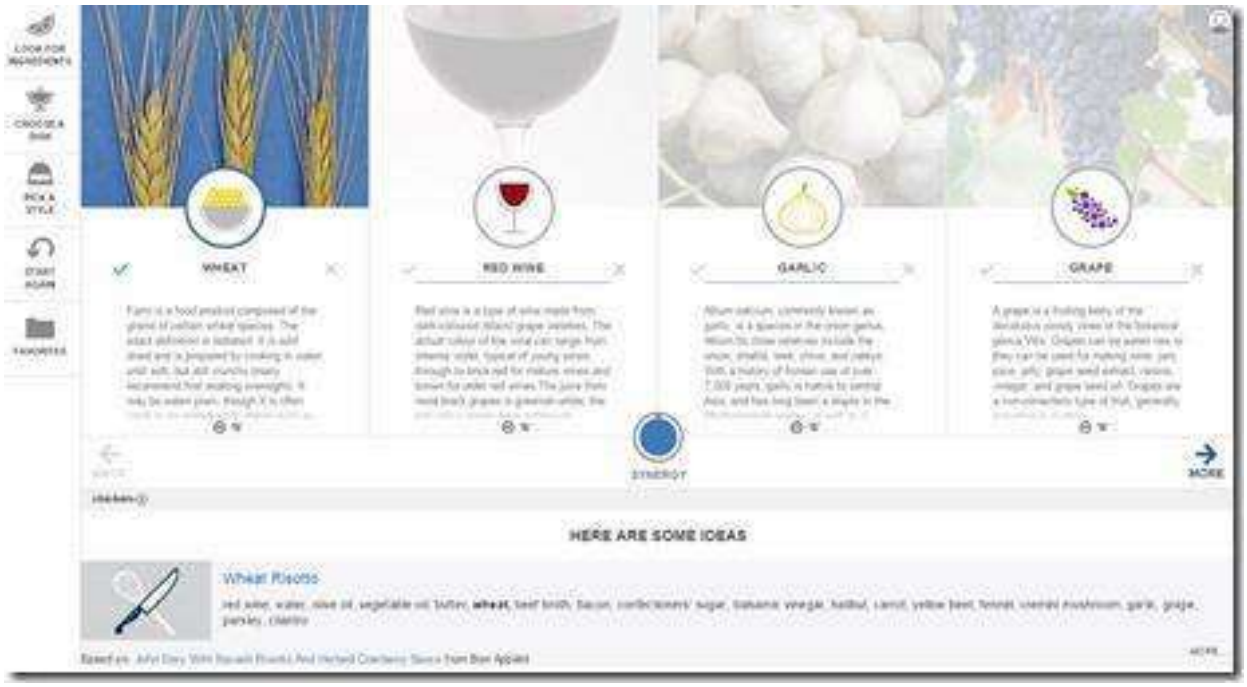
[14] Voir **ce numéro** de The American Lawyer qui en parle bien.

Dans les objets connectés, avec l'ouverture d'un **centre de recherche** dédié à l'IOT et Watson à Munich.

Dans les voyages avec la solution **Cognintive Travel** de la start-up **WayBlazer** qui est une sorte de concierge numérique commercialisé aux professionnels du tourisme. Une solution équivalente est proposée par**GoMoment**.

Thomson Reuters qui automatise la production d'études de marché. Peut-être sauront-ils générer d'autres courbes que les droits dans leurs prédictions et adopter les exponentielles et les gaussiennes !

Le spécialiste de l'équipement des centres d'appels **Genesys** va utiliser Watson pour améliorer ses services en analysant le flot de data généré par les appels clients.



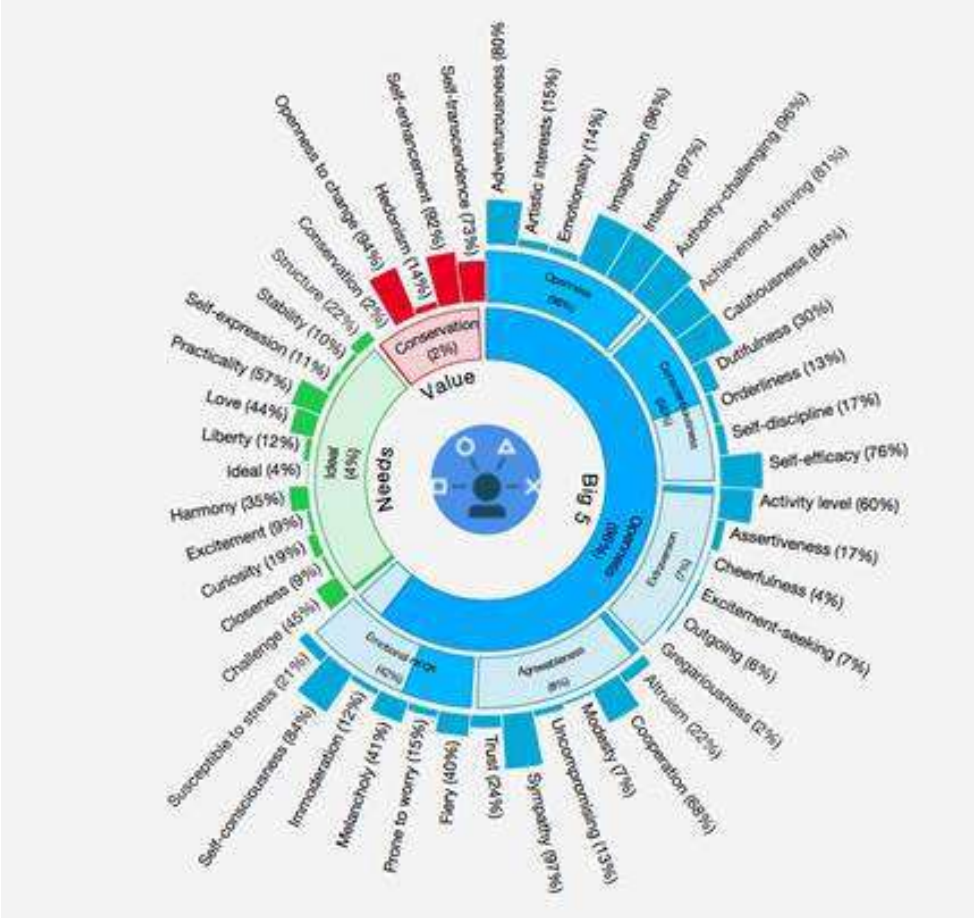
Le magazine US **Bon Appétit** utilise Watson pour proposer automatiquement des recettes de cuisine en fonction des ingrédients que l'on souhaite utiliser (*ci-dessus*). Il ne génère visiblement pas encore de recettes de cuisine automatiquement. Je rêve de mon côté d'une application (qui existe peut-être) qui pourrait me proposer des épices composées en fonction des épices dont je dispose dans ma cuisine !

La start-up de découverte de musique **Decibel** utilise aussi Watson dans son application MusicGeek pour faire de la recommandation.

IBM a fait l'acquisition de **Weather Company** pour $2B et ainsi alimenter Watson avec des données météorologiques destinées à diverses applications. Les données sont notamment utilisées par les assurances 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.

Watson est même utilisable pour **analyser votre personnalité** à partir de vos écrits ! Pour l'instant, la solution ne fonctionne qu'en anglais ce qui permettra aux bloggeurs et journalistes français d'éviter la création de classements à la noix ! La solution per-

met en tout cas de détecter l'humeur de l'auteur, comme **sa tristesse**. On voit bien Facebook, voir Périscope, utiliser ce genre de solution pour prévenir des suicides !



Gouvernements

Des études de cas de Watson dans la sécurité doivent bien exister sans être documentées, même si l'offre d'IBM n'est pas forcément ce que l'on peut trouver de mieux pour ces besoins particuliers.

On peut noter la campagne **Watson for president** qui vise, 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.

Il faut aussi tenir compte de la connaissance de l'Histoire qui influe sur les décisions des politiques. Le cerveau fonctionne très souvent par analogies. 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.

Pourrait-il répondre : 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 ? 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 du FN et des populismes dans le monde et la situation des années 1930 et d'avant seconde guerre mondiale ? Qu'est-ce qui est similaire et qu'est-ce qui est différent ? Comment anticiper la dimension émotionnelle qui fait bouger un peuple ? Quand est-ce que le peuple est au bord d'une révolte ? Comment l'anticiper ?

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, voir quand les deux parties sont irrationnelles avec ce même Saddam Hussein et Georges W. Bush en 2003 ?

Je m'étais aussi demandé **en 2013**, pour les 50 ans de l'assassinat de JFK, si un système de type Watson ne pourrait pas 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. 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 !

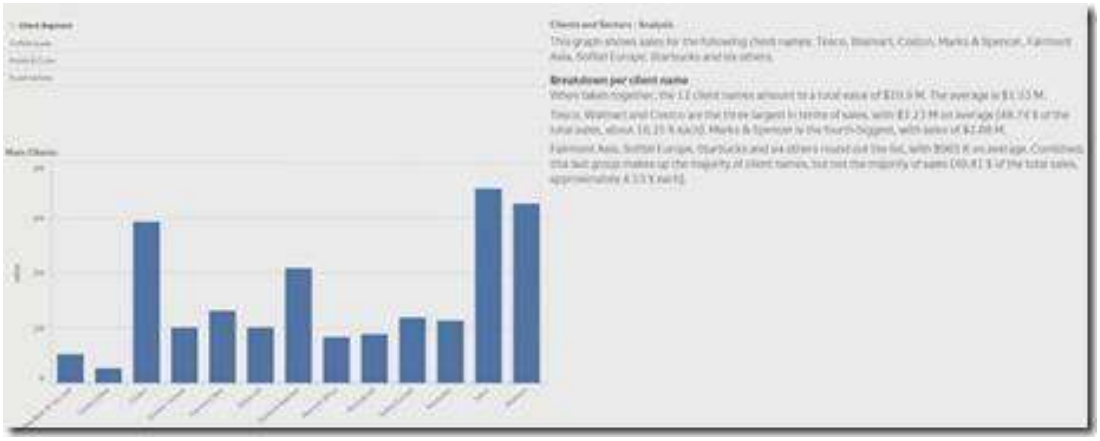
L'IA mis à toutes les sauces dans le marketing des start-ups

L'IA ou le machine learning deviennent des "selling points" de solutions logicielles et en particulier dans l'univers des start-ups. Ce sont l'équivalent des vrais morceaux de fruits dans les yaourts ! Sauf qu'on ne connait pas bien les fruits, leur origine et leur nature.

Rares sont les start-ups qui expliquent en détail les prouesses qu'elles réalisent grâce à l'IA et l'effort que cela leur a demandé. D'où viennent les logiciels ? Sont-ce des briques open source ou des briques spécifiques ? D'où viennent les données ? Nécessitent-elles de la préparation pour l'ingestion par le système ? Est-ce que le système a nécessité beaucoup de tests pour sa mise au point ? Ou au contraire, en faut-il beaucoup pour chaque projet client ?

Exemple de communication ambiguë de ces points de vue-là de la part d'**Yseop**, un éditeur de logiciels français spécialiste de la business intelligence et dont l'activité internationale se porte très bien :

*"A l'aide d'un puissant moteur d'intelligence artificielle créé par **Yseop**, Savvy rend les données compréhensibles pour tout le monde. Ainsi, le logiciel peut traduire une représentation graphique, quelle que soit son niveau de complexité, en un texte limpide, en anglais ou en français."*



En fait, Savvy est un plugin pour Excel qui traduit en texte compréhensible les données d'un graphe. Avec la **vidéo de démo**, on n'est pas plus avancé pour identifier les morceaux de fruits du yaourt d'IA du logiciel ! Il y a probablement un bon savoir faire derrière ce logiciel mais son label "IA" est des plus vague ! Cela donne un petit côté magique et mystérieux au logiciel ce qui lui donne probablement une certaines valeur, le sortant de la vase du logiciel de commodité.

J'ai pu obtenir une explication de la société sur le fonctionnement de son logiciel. Seulement voilà, il est difficile à traduire en langage naturel (moteur d'inférence d'ordre 2, …) ! Ca se complique sémantiquement quand des "algorithmes génétiques" sont utilisés. Ca y est, on exploite l'ADN de l'utilisateur ? Que nenni ! Il s'agit d'algorithmes de réseaux neuronaux qui combinent l'état de deux neurones pour conditionner celui de la suivante dans la chaîne de traitement, comme dans la combinaison des chromosomes dans la reproduction sexuée du vivant !

Chez **CrowdFlower**, on **enrichit** vos données grâce au machine learning ! Soit. Idem chez **Sift Science** qui lutte contre la fraude grâce encore au machine learning. On se retrouve à observer un marketing utilisant des mots magiques et où machine learning vient compléter l'arsenal du cloud et autre big data.

Chez **Groupe361**, on propose la solution Tanukis qui comprend "*le premier dispositif e-learning doté d'intelligence artificielle*" qui ressemble à un agent conversationnel, ajustant le parcours pédagogique en fonction du comportement de l'élève.

Il faudra un jour inventer des niveaux d'IA ou de machine learning. Histoire de pouvoir évaluer la taille des morceaux de fruits d'IA dans les yaourts à base d'IA ! Car le client et commentateur moyen a bien du mal à caractériser les morceaux de fruits et leur assemblage ainsi que les épaississants ! On peut aussi se demander si et quand les médias spécialisés seront en mesure de benchmarker les logiciels intégrant de l'IA dans leurs labos de tests.

Startups US de l'intelligence artificielle

Je vais essayer ici de segmenter ce marché, d'en identifier les tendances et lignes de force, de voir comment il se structure (verticalement, horizontalement) et comment il s'organise entre produits, données et services. Je vais aussi essayer d'identifier ce qui est rare dans ce marché.

Certains spécialistes de l'IA m'ont à juste titre fait remarquer que l'IA en était encore au stade artisanal et principalement de l'ordre 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 dans votre voiture haut de gamme semi-automatique.

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".

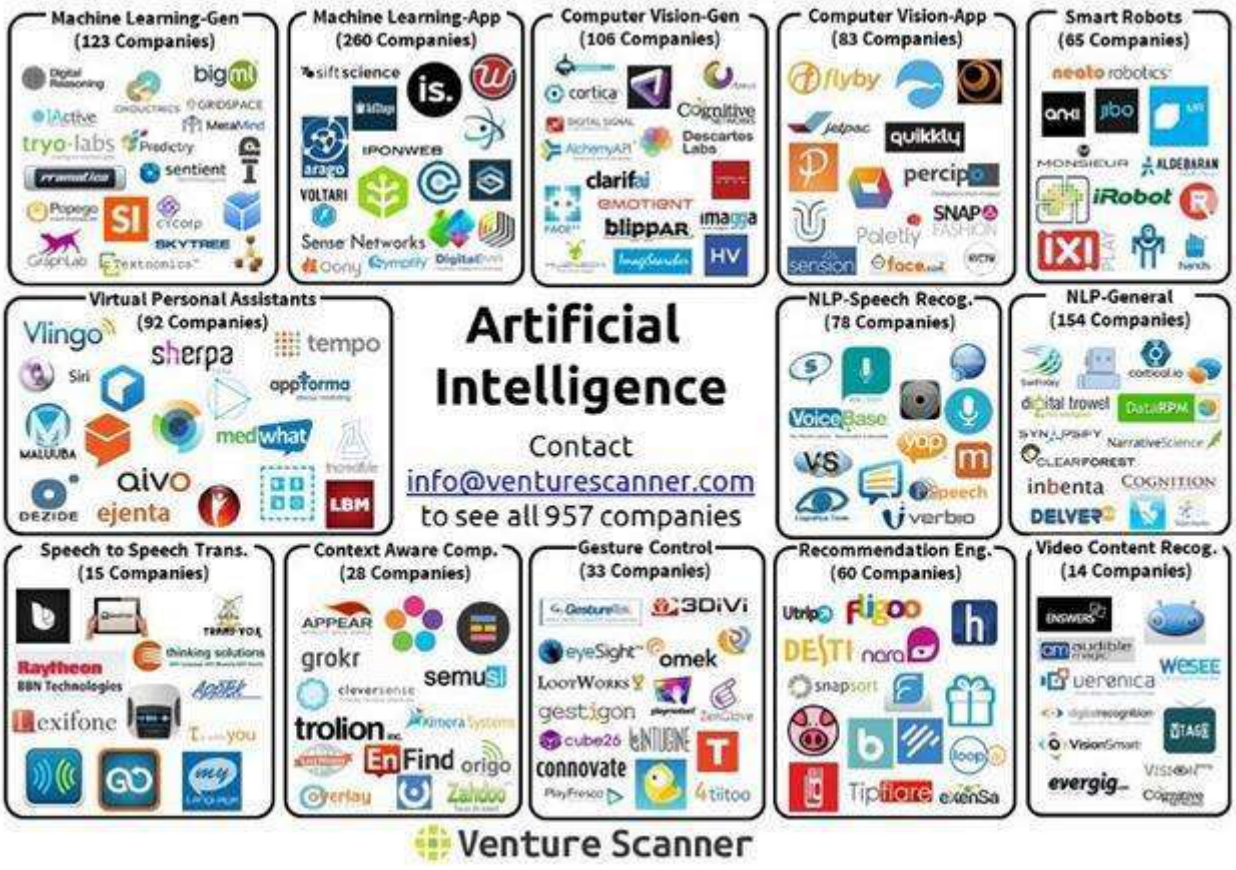
Cartographies des startups de l'intelligence artificielle

Pour cette partie, je vais m'appuyer sans vergogne sur ce suivi du secteur par le site **VentureScanner** qui était actualisé en mars 2016. Il organise le marché des startups de l'intelligence artificielle en 13 segments et évalue leur ancienneté et leur financement.

Au vu de ce grand schéma, j'organise cela avec pour commencer les segments du **Machine Learning** et du **Deep Learning** sont les plus représentés avec 123 startups identifiées pour les outils génériques et 260 pour des applications métier. On trouvait 684 startups utilisant du machine learning dans la **Crunchbase** en mars 2016.

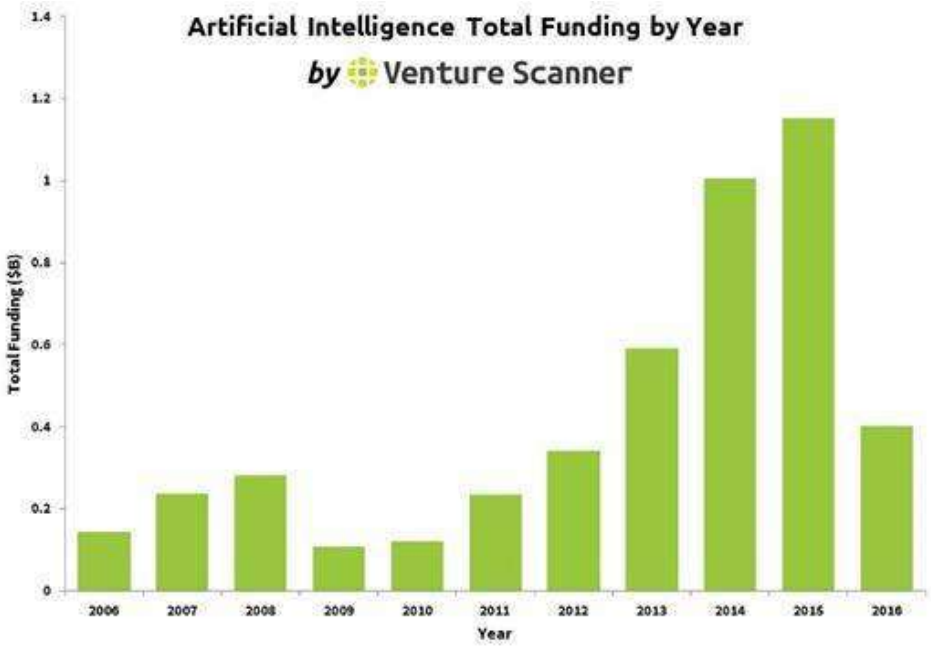
Ensuite, les **agents conversationnels** sont dans 92 startups, les **agents intelligents** qui comprennent leur environnement et agissent en conséquence sont 28. Suivent la **robotique** avec 65 startups, la **traduction automatique** avec 15 startups et les **moteurs de recommandation** qui représentent 66 startups.

Enfin, les startups gérant les solutions de perception : le **traitement du langage** avec 232 startups, la **vision artificielle** avec 189 startups, la **reconnaissance de vidéos** avec 14 startups et le **contrôle gestuel** avec 33 startups.

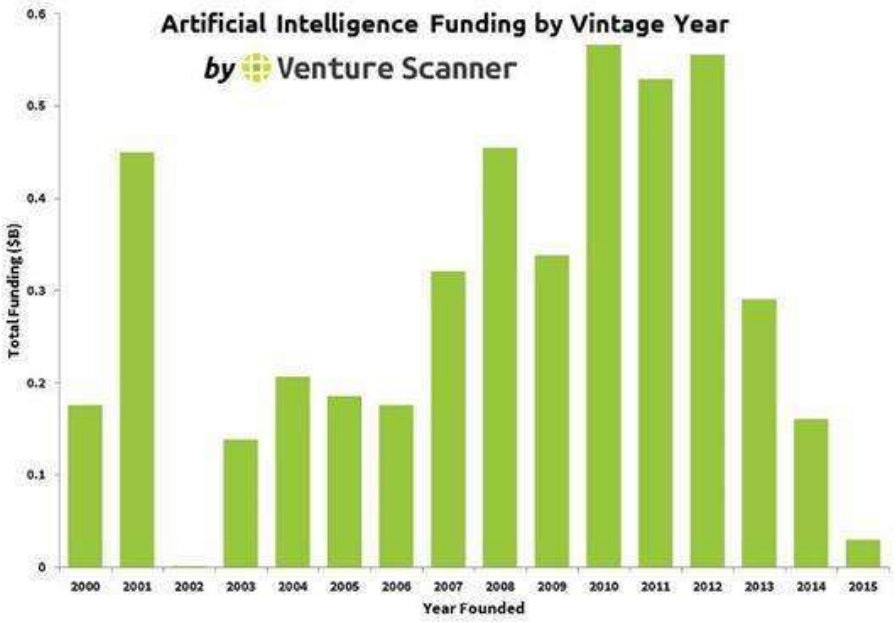


On constate une évolution à la hausse du financement des startups de ces secteurs. De 2009 à 2015, c'est une évolution constante, avec une courbe en cloche atteignant visiblement son pic en 2015.

Mais cette évolution a marqué de très nombreux secteurs d'activité comme les Fintechs, le retail, ou le e-commerce. L'IA n'est pas encore une priorité nette des VCs qui mettent encore le paquet sur des secteurs traditionnels. Nous avons ici $1,2B d'investissements dans l'IA pour $59B en 2015, en tout rien qu'aux USA !



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 entre 2010 et 2012 mais un gros volume de startups créées entre 2006 et 2010. Elles sont encore là car elles doivent probablement cibler des marchés d'entreprises. Les investisseurs ont tendance à financer des startups plutôt matures dans ce secteur. Les startups les plus anciennes de l'IA sont celles de la reconnaissance de la parole et de la vidéo, qui ont respectivement 8 et 6,5 ans d'ancienneté.



Ces startups ont généralement quelques points communs marquants :

- Elles ont quasiment toutes 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.
- On y trouve régulièrement les ombres de la **DARPA, de la NSA et de la CIA** comme clients voire même comme investisseurs pour cette dernière ! 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 !
- 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".

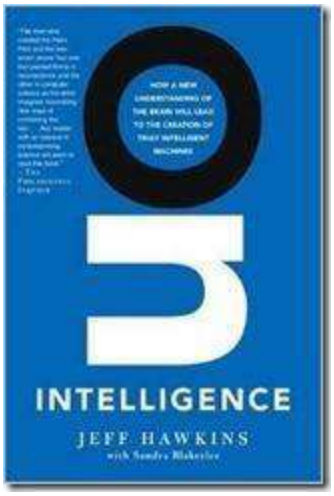
- Les solutions sont très souvent proposées sous la forme **d'APIs en cloud** mais les approches plateformes sont encore en devenir et balbusentienttiantes car elles ne bénéficient pas d'un effet push/pull courant dans le grand public (la demande pour des smartphones Android entrainant celles d'applications tournant dessus).
- Les **levées de fonds** sont encore relativement modestes dans l'ensemble. On dépasse dans de rares cas les $100m. Ce n'est pas beaucoup par rapport à plus de $1B réalisées par des licornes telles que Pinterest, où l'intensité technologique est plus faible. Les licornes sont presque toutes des startups grand public.
- Le secteur donne lieu à de nombreuses **acquisitions** mais Google n'a pas acquis tout ce qui était intéressant ! Autant on sait que Google a bien ratissé le secteur par de nombreuses acquisitions telles que celle de DeepMind (UK) en 2014, autant on peut constater que nombreuses aussi sont les startups créées par d'anciens de Google et notamment des Google Labs.
- 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.

Dans ce qui va suivre, je vais indiquer la date de création des startups ainsi que les montants levés entre parenthèses lorsqu'ils sont disponibles. Même si les montants levés ne sont pas une indication suffisante de succès, elles montrent que la société a au moins attiré le regard et l'argent d'investisseurs. Les financements qui dépassent les $20m indiquent une "traction" qui peut avoir un impact mondial assez rapidement.

Deep Learning et Machine Learning

C'est la catégorie de startups la plus 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, NC) 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 (**PDF gratuit**).



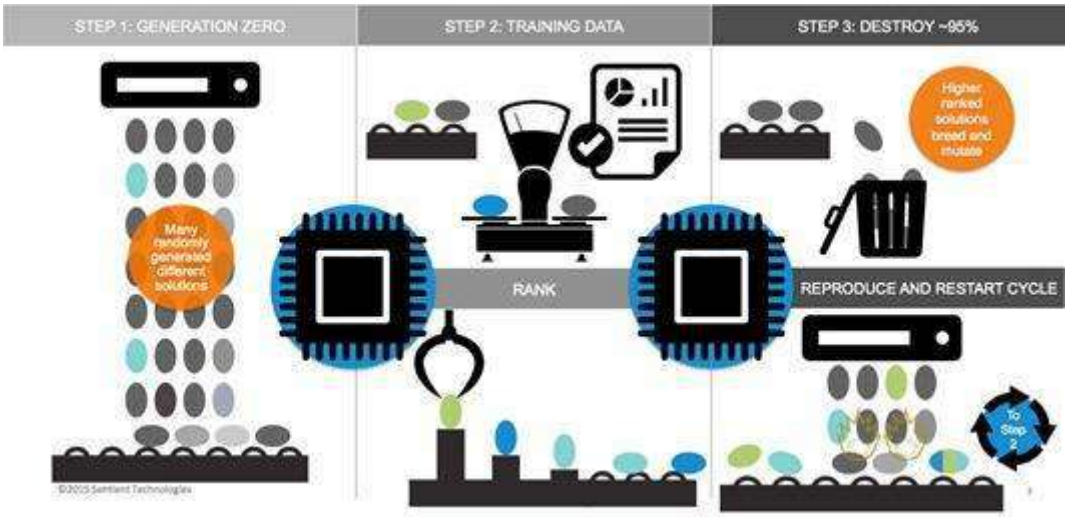
Hawkins pense que le cerveau est principalement une machine prédictive qui n'est pas forcément dotée d'une capacité de calcul parallèle intensive mais plutôt d'une mémoire associative rapidement accessible. Il insiste sur l'importance du temps dans les mécanismes de rétropropagation mise en œuvre dans les réseaux neuronaux uniquement dans les phases d'apprentissage. Alors que le cerveau bénéficie d'une mise à jour sensorielle permanente.

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 (voir ces critiques chez **Jeff Kramer**, **Ben Goertzel** et sur **Quora**).

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, $135m, dernier tour de financement en 2014) 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.



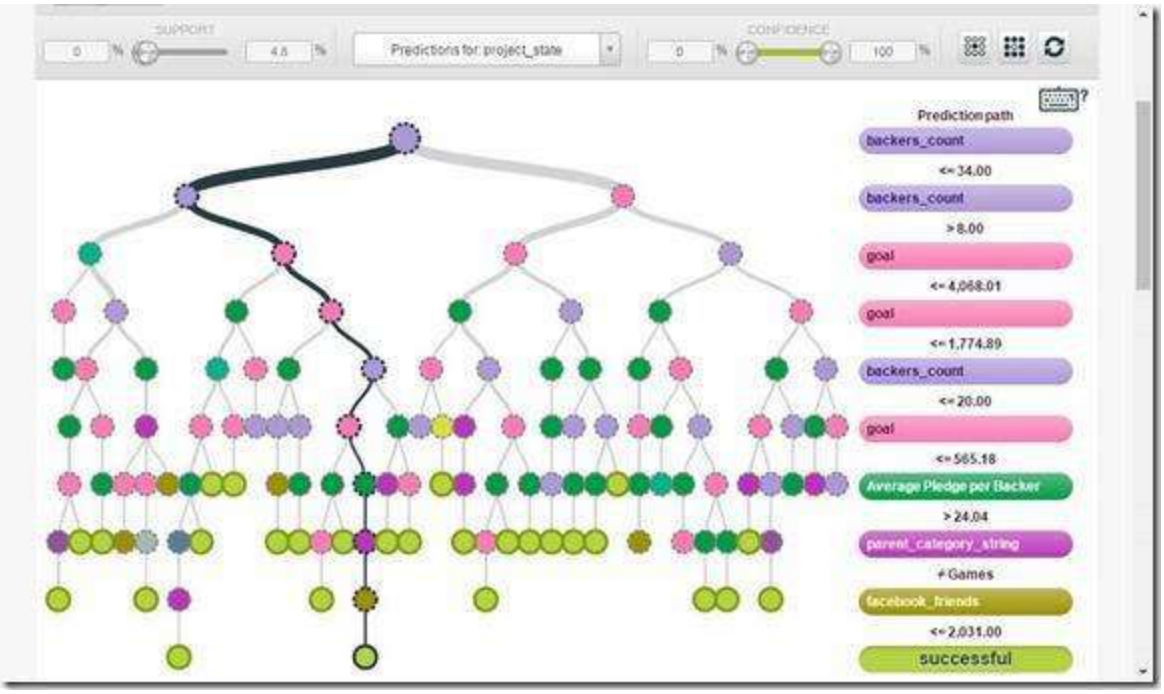
Digital Reasoning (2000, $52m, dernière levée en 2016) a été créée par des anciens d'Oracle et de la CIA (entre autres provenances) et est financée par In-Q-Tel, le fonds d'investissement de cette dernière. Sa solution d'analyse de données est utilisée par le renseignement et la défense US ainsi que 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).

Metamind (2014, $14m) fait de la classification automatique d'images et de textes, un peu comme ce que propose IBM Watson pour la partie texte. Elle a été créé par une équipe d'anciens de Stanford.

Scaled Inference (2014, $13,6m) 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 prédiction. Startup créé par un ancien de Google. Solution pas encore disponible.

Skymind (2014, NC) 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, NC) 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, $98m) 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, $29,4m) 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.

Workfusion (2010, $36,3m) propose une solution en cloud d'orchestration et de consolidation de données pour les entreprises. Elle s'appuie sur de l'apprentissage supervisé d'outils traitant de gros volumes de données par des travailleurs crowdsourcés, dans divers métiers comme les services financiers, la comptabilité et le e-commerce. Le projet est issu de travaux de recherche du MIT.

Nutonian (2011, $4m) propose une solution d'extraction de données intelligente, capable d'identifier des tendances cachées dans les données.

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.

Context Relevant (2012, $44m) 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 ($36m) 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.

Skytree (2012, $20,5m) propose une autre solution de moteur de prédiction, Skyree Infinity qui peut par exemple prédire le comportement des consommateurs et identifier des segments d'acheteurs potentiels de produits précis. La startup propose SkyTree Express en téléchargement gratuit pour analyser jusqu'à 100 millions d'éléments. Ils sont financés par la CIA via son fonds d'investissement In-Q-Tel en plus de Samsung.

Sentenai (2015, $1,8m) propose aussi 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, $23m) et **Lavastorm** (1999, $10m).

IA pour la recherche visuelle

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.

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%.

Clarifai (2013, $72m) propose une API en cloud permettant d'accéder à leurs fonctions de reconnaissance d'images.

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.

Camio (2013) fournit une solution en cloud d'exploitation de vidéos de caméras de surveillance.

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 !

Descartes Labs (2014, $8,28m) exploite via deep learning les données d'image satellite pour y découvrir comment évolue la production agricole, le cadastre des villes ou autres données géographiques.

En complément de ces startups, on trouve aussi des startups spécialisées dans le traitement du langage. Là encore, tout n'est pas chez les GAFA ou chez Nuance. On peut notamment citer **DefinedCrowd** (**vidéo** avec son ukulélé de circonstance), **Weotta** (2011) et **MindMeld** (2014).

Robots conversationnels

Les robots conversationnels ou "chat bots" sont très en vogue en ce moment. On les retrouve dans de nombreuses startups ainsi que dans diverses offres de grandes entreprises du numérique (Facebook, Google, …).

Toutes visent à automatiser le service client en ligne dans les sites de e-commerce ou autres, visant la réussite du fameux test de Turing qui définit une intelligence artificielle comme une intelligence indistinctible de celle de l'homme dans de telles discussions par le biais d'échanges textuels.

Dans les startups, on peut notamment citer…

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 très précis, comme l'offre d'une société particulière.

Msg.ai (2014, $2,7m) est une startup qui propose des chatbots pour les sites de vente en ligne. Elle est notamment déployée chez Sony.

Niki.ai (NC) est une startup indienne qui, comme la précédente, propose des chatbots pour les sites de vente en ligne, notamment dans les services (transports, voyage aérien, santé).

ReplyYes (2015, $3,5m) est une autre startup américaine, de Seattle, qui propose une solution de chatbots pour les sites de vente en ligne associant machine learning et opérateurs humains. Ils ont deux spinoffs, l'une qui vend des disques vinyles (The Edit) et l'autres, des BD (Origin Bound). The Edit aurait vendu $1m de vinyles en huit mois.

Semantic Machines (2015, $12,38) est une startup de Boston et Berkeley qui propose des chatbots pouvant être intégré 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.

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. 385 000 chatbots ont été générés avec (mai 2016). Ils sont intégrables dans divers environnements de messagerie instantanés tels que Slack et Whatsapp.

TARA (NC) est une startup de San Francisco qui propose un robot conversationnel de gestion du recrutement de freelances.

Ces solutions de chatbots ne sont pas évidentes à départager. Elles reposent probablement sur des techniques voisines et se distinguent parfois plus par les marchés visés que par leur performance absolue.

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.

Applications sectorielles du machine learning

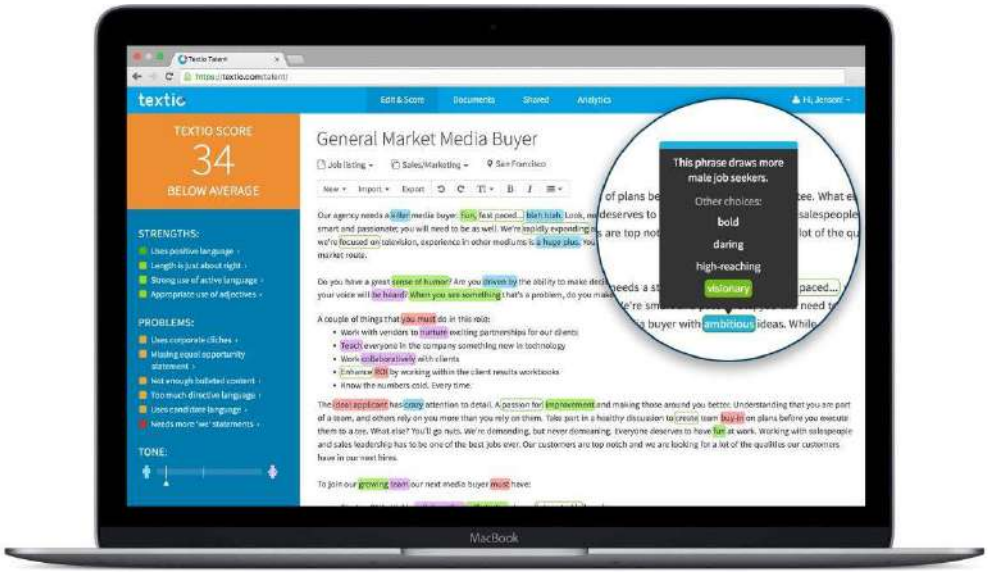
L'autre plus grand groupe de startups de l'IA couvre celles qui utilisent les techniques de machine learning et deep learning, le plus souvent de nature non précisée, et qui ciblent des marchés spécifiques. Je vous les survole très rapidement, histoire de se faire une idée des applications les plus courantes.

Dans les **services financiers** : avec de l'optimisation de taux d'intérêts de prêts (**LendUp**, qui a levé $150m), pour la détection de fraude (**Sift Science**, $23,6m, **Riskified**, $30m), pour le credit rating d'emprunteurs basée sur les réseaux sociaux (**TrustingSocial**) dans l'optimisation de planification financière (**Anaplan**, $234m, **Adaptive Planning**, $22,5m, **Trufa**, $10,9m), pour l'optimisation d'investissements (**DataFox**, $7,7m) et même pour identifier des startups dans lesquelles investir (**Mattermark**, $17,2m). Il y a aussi **Vanare**,**Sens.ai** et **WealthArc**.

Dans le **commerce** : prédiction du trafic dans les magasins (**Percolata**, $5m), l'optimisation du parcours client en ligne (**Gainsight**, $104m, **Jetlore,** $7m, **OnCorps**, $2,3m)**,** pour trouver la bonne taille et pour les hommes (**Thread**), pour optimiser l'activité de commerciaux et prévoir le comportement des clients (**InsideSales**, $201m !, **Gainsight**, $104m, **Lattice**, $64,7m, **Clari**, $26m, **Wise.io**, $2,61m, **Spiro**, $1,5m). On peut ajouter**Dato** (2013, $23,5m) qui propose un système de recommandation dans le ecommerce basé sur du machine learning. La startup a été montée par des anciens de Carnegie Mellon sous la forme initiale d'un projet open source.

Dans le **marketing**, pour l'optimisation des messages et contenus (**Captora**, $27m, **Persado,** $36m), pour la gestion ou l'analyse des données issues des médias sociaux (**Meshfire**, $350K, **Cortex**, $500K,**SimpleReach**, $10,6m).

Dans les **ressources humaines** avec de l'analyse prédictive pour identifier des talents à chasser (**Entelo**, $8,7m, **Gild**, $26mn). Il y a aussi **Textio** qui aide à rédiger des annonces d'emploi efficaces et analyse les réponses des candidats (*exemple ci-dessous*).



Dans les **services juridiques** avec divers systèmes d'interrogation de bases de connaissances (**Casetext**, $8,8m, **Judicata**, $7,8m, et **Ross Intelligence**, qui s'appuie sur IBM Watson). Il y a aussi **Kira** et **Legal Robot**.

Dans les **universités** pour les aider à recruter les meilleurs étudiants (**Plexuss**) ou, au contraire, pour aider ces derniers à trouver la meilleure université (**Admitster**).

Dans la **recherche scientifique** : pour gérer une communauté mondiale de data scientists (**Kaggle**, $12,75m), pour découvrir des étoiles selon leurs caractéristiques ().

Dans la **sécurité informatique** pour détecter les tentatives de phishing (**GreatHorn**, $2,6m) ou avec **Lookout** (2007, $282m) qui sécurise les mobiles avec un modèle prédictif.

Dans l'**agriculture** pour robotiser la culture et personnaliser le soin de chaque plan (**Blue River Technology**, $30,35m). La société propose 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*).

Dans les **contenus** avec le canadien **Landr** ($10,4m) et son service en cloud pour automatiser le mixage audio et créer des morceaux de musique agréables à l'écoute. Il y a aussi **Narrative Science** (2010, $29,4m) qui est capable de rédiger tout seul des textes à partir de données structurées et non structurées, utilisé notamment dans les médias et le marketing.

Applications dans la santé

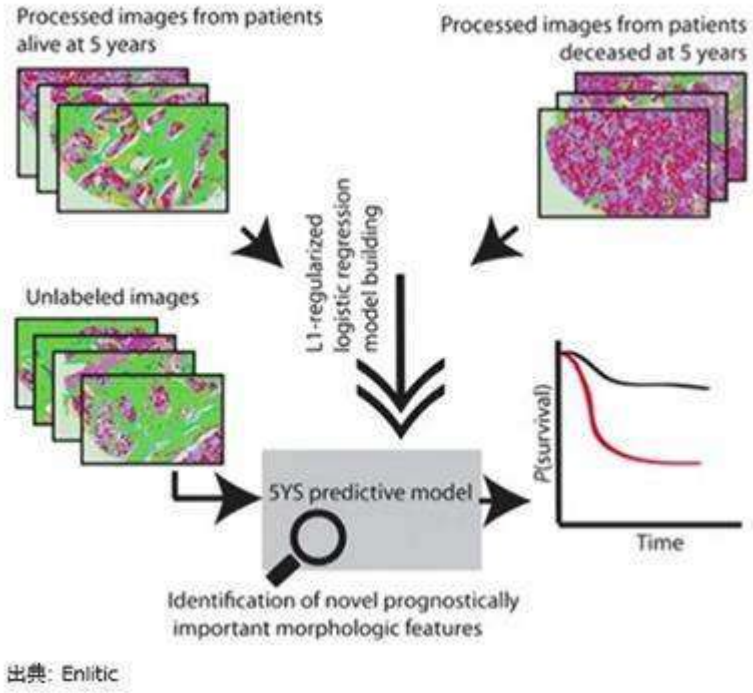
C'est le marché vertical le plus attirant pour les startups de l'IA avec celui de la finance et du commerce. L'IA est notamment utilisée dans la génomique et dans l'aide au diagnostic dans la lignée de la solution en oncologie que nous avons vue au sujet d'IBM Watson. Ce qui suit n'est probablement qu'une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

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. On appelle cela les "genome-wide association study" (GWAS) qui font des analyses de corrélations entre modifications des gênes et pathologies (le "phénotype"). Les analyses réalisées par Deep Genomics ont la particularité d'intégrer tout le cycle de vie des gênes et notamment leur épissage – qui correspond à l'extraction de la partie codante des gênes – jusqu'à leur translation, à savoir la conversion de l'ARN qui résulte de l'épissage en protéines dans les ribosomes. Ils proposent en open source leur base de données SPIDEX de mutations de gênes et de leurs effets sur leur épissage[15]. 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 renouveau dans les réseaux neuronaux au milieu des années 2000 et qui est maintenant chez Google.

Enlitic (2014, $15m) propose de l'aide au diagnostic en s'appuyant principalement sur résultats de systèmes d'imagerie médical (IRM, scanner, radios) et sur du deep learning. C'est une sorte d'équivalent apparemment généraliste d'IBM Watson qui se positionne plutôt dans la prévention, détectant des pathologies émergentes le plus tôt possible, notamment les cancers du poumon. Il aide aussi à identifier plusieurs pathologies simultanément[16].

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

[16] 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.



出典: Enlitic

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 auto-traitement de certaines pathologies par les patients.

Lumiata (2013, $10m) 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 !

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.

Cognitive Scale (2013) a créé la solution Cognitive Clouds. Elle est notamment 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.

atomwise (2012, $6,35m) 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.

MedAware (2012, $1m) 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 !

Startups acquises par les grands du numérique

Après la partie précédente, couvrons quelques-unes des startups manquantes du secteur, acquises par de grandes entreprises et notamment par les GAFA et autre IBM, Microsoft, Oracle, LinkedIn et Salesforce.

C'est le lot commun de vagues d'innovations que de générer ce genre d'acquisitions. On en déduit que les grandes entreprises ont le souci de s'adapter rapidement aux évolutions du marché et de ne pas rater un train qui part ou qui est déjà en marche. L'IA n'y échappe donc pas. Cela a commencé il y a pas mal d'années avec des acquisitions portant d'abord sur la recherche, le traitement de l'image et la reconnaissance de la parole. Les acquisitions les plus récentes portent plutôt sur le machine learning adapté au big data.

Voici donc un tour d'horizon de ces principales acquisitions et, au passage, des stratégies d'IA associées pour ces grands groupes du numérique.

Google

L'actualité abonde depuis 2014 d'acquisitions médiatisées de startups de l'IA par ces grands acteurs du numérique. Cela alimente quelques fantasmes sur leurs avancées qui sont quelque peu enjolivées. Elles sont notamment focalisées sur Google qui aurait, selon les commentateurs, acquis tout ce qui existait de bien autour de 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. Ce qui est rare et cher. En 2014, une **estimation** portait à une cinquantaine le nombre de chercheurs en machine learning dans le monde. A vrai dire, il y en avait certainement beaucoup plus. On ne compte pas les PhD qui font ou ont fait de la re-

cherche en machine learning. Mais les "bons" sont probablement peu nombreux comme dans toutes les disciplines de la recherche.

Google n'en était pas à son premier coup. Ils avaient auparavant mis la main sur la société de reconnaissance vocale **SayNow** en 2011, qui se retrouve probablement dans la commande OK Google sur mobiles Android. Puis sur **Viewdle** et de **PittPatt** en 2012, qui faisaient tous les deux de la reconnaissance faciale et de mouvements. Ensuite, en 2013, sur le spécialiste des réseaux neuronaux **Dnnresearch**, avec le canadien Geoff Hinton au passage, père du renouveau des réseaux neuronaux, évoqué dans le **second article** de cette série.

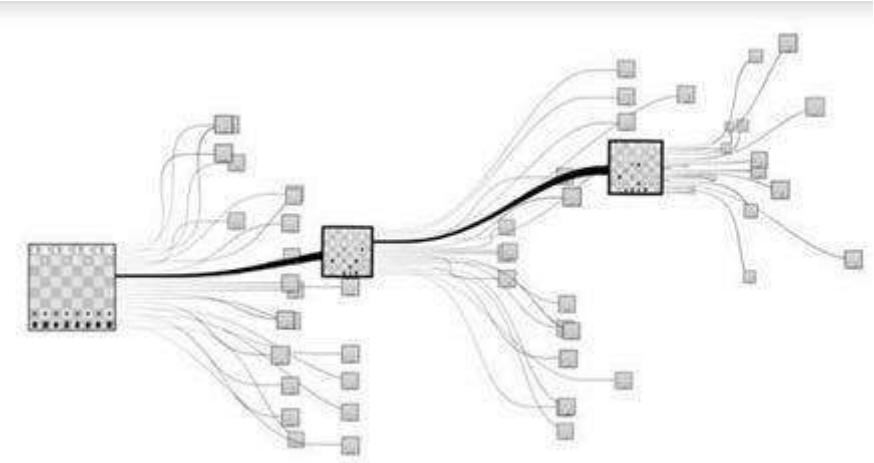
Ont suivi **Dark Blue Labs** et **Vision Factory**, deux sociétés d'Oxford qui n'ont pas levé de fonds. S'y ajoutèrent la solution de recommandation d'images **JetPack**, le spécialiste de la traduction automatique **Quest Visual**, et celui de la reconnaissance de mouvements **Flutter** qui a probablement 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 palanquée de startups dans la robotique avec **Schaft** (robot humanoïde et bras articulé, japonais), **Industrial Perception** (robots industriels, spécialisé dans la vision 3D), **Redwood Robotics** (bras robotisés, issue du SRI et acquise un an après sa création), **Meka Robotics** (aussi dans les bras robotisés, qui avait contribué à la création de Redwood Robotics), **Holomni** (roues robotisées), **Bot & Dolly** (bras articulés à mouvements très souples servant aux tournages de cinéma), **Autofuss** (encore des bras articulés) et surtout **Boston Dynamics**, connu pour ses robots médiatisés doués de capacité de marche à quatre puis deux pattes mais que Google est **en train de céder**.

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 peu d'applications commerciales pour l'instant car ces technologies sont toujours en phase de gestation ou destinées à des marchés de niche. 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**. 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 avec la victoire de la solution AlphaGo 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**). AlphaGo avait déjà battu le champion européen, Fan Jui, en octobre 2015. 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 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" (*représenté graphiquement ci-dessus*) ou MCTS et exploiter une base de jeux permettant d'identifier des tactiques gagnantes. A ce jeu-là, cependant, AlphaGo **ne gagne pas systématiquement**. Mais c'est une bonne pub pour DeepMind dont les solutions de machine learning ont d'autres applications comme la **curation de médias** ou **la santé**. 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 2016, DeepMind était toujours un laboratoire de recherche. Ils planchent sur **DQN**, un réseau neuronal profond doté de capacités d'auto-apprentissage et **DeepMind Health**, qui donne lieu à une collaboration avec la NHS britannique et dans l'application Streams de détection de blessures aux reins dans les urgences.

Le français Yann LeCun, maintenant chez Facebook, a **trouvé de son côté** que le progrès accompli par AlphaGo n'est **pas révolutionnaire** pour autant, surtout en termes d'intelligence, un mot valise dont la définition est à géométrie variable. Il met en évidence le rôle de l'architecture matérielle qui est mise en œuvre (280 GPUs et 1920 CPUs, même si ce n'est pas forcément celle qui l'était dans les matchs historiques).

Contrairement aux nombreuses startups évoquées dans la partie précédente, 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). Il s'approvisionne en technologies et compétences de manière assez classique, pas du tout dans un modèle d'innovation ouverte. L'enjeu est de transformer ces avancées en plateformes, comme il commence déjà à le faire en publiant des **APIs d'IA dans le cloud** pour les développeurs, dont fait partie TensorFlow, une bibliothèque open source de machine learning.

Comme ce fut le cas dans le Web 2.0, la position des acteurs est à la fois technologique mais aussi liée à la capacité à satisfaire des besoins importants d'utilisateurs, à créer des écosystèmes applicatifs et à trouver un bon modèle économique. Google n'y est pas toujours arrivé, comme avec Google Reader (abandonné), Picasa (abandonné au profit de Google Photos), Google Wave (transformé), Google+ (mal en point) ou **Google Compare** (récemment abandonné).

L'IA est en tout cas un domaine suffisamment ouvert pour que l'on ne se contente pas de ne prêter qu'aux riches !

D'ailleurs, petite question : quels travaux de recherche en IA a pu créer et publier Ray Kurweil depuis qu'il est chez Google, à savoir 2012 ? Il planche en théorie sur le traitement du langage naturel et sur la création d'une AGI (intelligence artificielle généraliste). On n'en sait pas plus ! Avant d'être chez Google, Kurzweil était à la fois un serial-inventeur et un serial-entrepreneur en plus d'être auteur de nombreux livres, donc plutôt prolifique. Depuis, c'est un grand silence. Les exponentielles de progrès ne sont pas universelles !

IBM

IBM est le second acteur auquel les médias font le plus souvent référence concernant l'IA, notamment depuis qu'il communique fort bien sur les performances et applications diverses de Watson.

IBM est aussi coutumier des acquisitions, qui servent surtout à enrichir son offre de logiciels d'entreprises, source de profits de même niveau que ses activités de services pourtant trois fois plus importantes en chiffre d'affaire. L'offre logicielle d'IBM s'est

agrandie depuis les années 1990 par le fait de nombreuses acquisitions avec Lotus, Rational, Tivoli pour ne nommer que les plus connues. Et cela a continué depuis.

IBM a investi au moins $7B en acquisitions dans l'IA, bien plus que Google ne l'a fait. Il a notamment absorbé en 2014 la startup **Cognea**, créatrice d'un agent conversationnel, **AlchemyAPI** (2005), une startup de deep learning d'analyse de textes et d'images, de reconnaissance de visages, de tagging automatique d'images acquise en 2015,**IRIS Analytics**, une startup allemande d'analyse temps-réel dédiée à la détection de fraudes aux moyens de paiement, s'appuyant sur du machine learning, **Explorys** et la solution de gestion de données patients **Phytel**, intégrés dans les solutions santé de Watson, acquises en 2016.

Au-delà de l'acquisition de technologies et d'équipes, IBM s'intéresse aux données permettant d'alimenter Watson de manière générique. Cela s'est manifesté avec l'acquisition des activités données de **The Weather Channel** en 2015 permettant de créer une base météorologique conséquente utile dans différents marchés (tourisme, agriculture, énergie, transports) ainsi que de **Truven Health Analytics** pour $2,6B, qui gérait les données – probablement anonymisées – sur le cout et les traitements de 200 millions de patients.

Cette démarche est originale. On ne l'observe pas chez Google ou les autres acteurs évoqués ici. Les données de certains domaines sont en effet plus rares que les algorithmes de machine et deep learning et de réseaux neuronaux qui sont plutôt monnaie courante. Le couplage données + big data + machine learning auto-apprenant permet de créer des bases uniques. D'où leur valeur économique pour IBM. CQFD.

Microsoft

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. Les équipes de Microsoft Research sont à l'origine d'avancées comme le système de dialogue en langage naturel Cortana.

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**).

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 !

Les acquisitions dans les startups de l'IA sont cependant peu nombreuses chez Microsoft. On peut surtout citer **Revolution Analytics**, qui faisait de l'analyse prédictive s'appuyant sur le langage open source R, acquise récemment en 2016. Un moyen de s'attirer un écosystème de développeurs ! Et puis, début 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.

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 entre le 28 mars et le 1ier avril 2016 (voir les vidéos de keynotes du **premier jour** et du **second jour**). C'est tout frais ! Cf ce **bon résumé** de Build 2016 dans l'Usine Digitale.

C'est dans ce cadre que Microsoft propose 22 APIs dans ses Cognitive Services (anciennement Projet Oxford) associés à l'agent Cortana Intelligence Suite. Ce sont des APIs sensorielles qui font de la reconnaissance d'images, audio, du traitement du langage naturel (NLP). L'agent intègre les services en cloud Azure Machine Learning, lancés en 2015. Microsoft a notamment démontré son CRIS (Custom recognition intelligence service), un outil de "speech to text" capable d'interpréter les paroles approximatives d'enfant en bas âge. Microsoft propose aussi en open source son son framework de "deep learning" CNTK (Computational Network Toolkit) depuis fin 2015.

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 publie *aucun* papier dans le domaine de l'IA. C'est tout le contraire de l'innovation ouverte !

Les acquisitions d'Apple sont peu nombreuses en règle général. Dans l'IA, on peut compter **Emotient** pour la reconnaissance des visages et des émotions, **VocalIQ**, qui devait enrichir les fonctionnalités de reconnaissance de la parole de SIRI en ajoutant de l'auto-apprentissage, ainsi que **Perceptio**, 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**, 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 comble les trous dans l'IA via son partenariat avec IBM qui porte notamment sur Watson. Il est cependant probable qu'Apple devra faire quelques acquisitions dans le cadre de son projet de voiture automatique.

Facebook

Facebook a fait parler de lui côté IA en lançant le FAIR (Facebook Artifical Intelligence Research) qui est dirigé par le chercheur français **Yann LeCun** et qui est ins-

tallé à Paris depuis 2015. Yann LeCun est l'un des pères du machine learning. Il est aussi depuis peu professeur au Collège de France sur le deep learning[17].

Les équipes de Facebook planchent sur la reconnaissance automatique de sports dans les vidéos ou de chiens dans les photos. Facebook vient aussi de démontrer une fonction qui décrit le contenu de photos, adapté aux aveugles, presque simultanément à une fonction du même genre proposée aux aveugles par Microsoft.

Facebook a aussi fait l'acquisition de **Wit.ai**, 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 !

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 !

Facebook lançait à sa dernière conférence développeurs en avril 2016 sa solution d'APIs pour le développement de chatbots, son Bot Framework.

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).

[17] Voir sa **leçon inaugurale** qui fait un très bon panorama technique du machine learning.

Autres grands acteurs du numérique

Dans les acquisitions de startups de l'IA par d'autres grands acteurs du numérique, on peut citer quelques mouvements récents, même si on est loin du tsunami. Pourquoi donc ? Est-ce lié aux domaines d'activités de ces entreprises ?

- La startup **Connectifier**, spécialisée dans l'identification de profils pour le recrutement, acquise par LinkedIn début 2016. LinkedIn a aussi acquis **Bright** en 2014, un spécialiste du recrutement basé sur de la recommandation et s'appuyant sur du machine learning pour l'analyse sémantique des CVs.
- Le moteur de recherche visuel mobile **SnapTell** a été acquis par Amazon en 2009, l'une des rares acquisitions de ce dernier semblant intégrer des solutions ou technologies d'IA avec le spécialiste de la reconnaissance d'images **Orbeus**, acquis en avril 2016. Ce qui peut vouloir dire que la solution en cloud **Amazon Machine Learning**, lancée en 2015, est d'origine interne. Elle permet de réaliser des recommandations d'achats, comme dans le service d'Amazon. Il pourrait en être de même d'Alexa, le moteur de reconnaissance vocale d'Amazon qui est notammet intégré à Amazon Echo et est aussi exploitable via ses APIs par des développeurs tiers.
- Le spécialiste du machine learning **PredictionIO** acquis par SalesForce également début 2016, ainsi que **Metamind** début avril 2016. Je les avais couverts rapidement dans l'**article précédent** alors qu'ils étaient encore indépendants !
- L'application de recommandation mobile **Livestar** acquise par Pinterest en 2013.
- La startup de machine learning et d'analytics **Whetlab** acquise par Twitter en 2015.
- Le spécialiste de data analytics **BigMachines** acquis par Oracle en 2013 pour $300m.

A chaque fois, il s'agit visiblement d'applications métiers et pas forcément de startups qui ont réellement fait avancer l'état de l'art de l'IA.

Il sera intéressant d'observer les acquisitions de startups d'ici fin 2016 pour identifier une éventuelle croissance des catégories IA et machine learning. Une liste à jour est disponible **sur CBInsights**.

Startups françaises de l'intelligence artificielle

Après avoir fait le tour des stratégies d'IA de quelques grands acteurs du numérique dont Google, IBM, Microsoft et Facebook et à leurs acquisitions, revenons aux startups du secteur en nous intéressant aux françaises.

Il est clair que l'IA est l'une des technologies clés du numérique, aujourd'hui et demain. Donc, au lieu de chercher à créer un Google, un Facebook ou un système d'exploitation français, il serait bon de s'intéresser à ce domaine prometteur, surtout dans la mesure où les plateformes correspondantes sont encore en devenir.

La recherche en IA en France

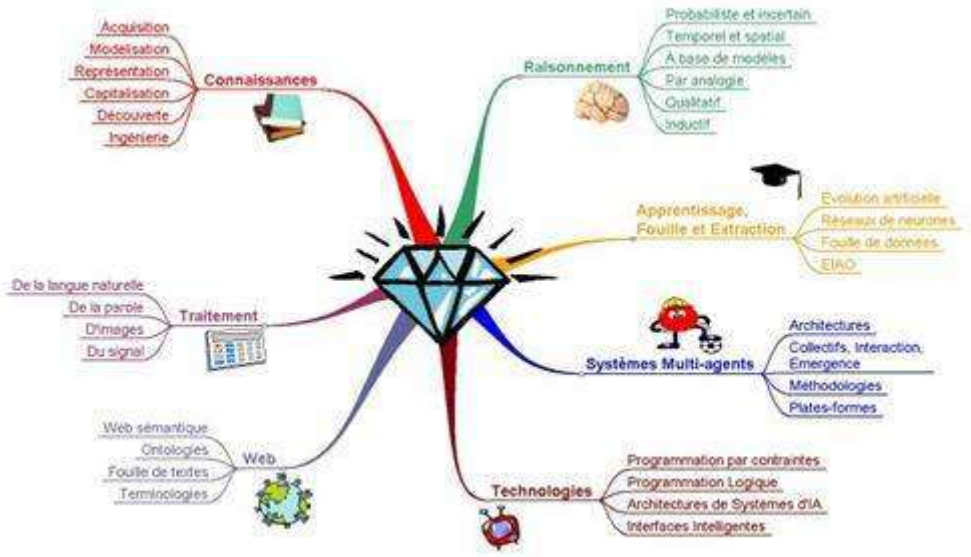
La recherche en IA est disséminée dans plusieurs laboratoires et dans des projets collaboratifs associant laboratoires publics et universités. Les deux premiers organismes se focalisant sur l'IA sont l'INRIA et le CNRS.

Que fait l'INRIA ? Un grand nombre des projets de recherche fondamentale en IA référencés sur leur site font appel aux techniques de l'IA, même s'ils ne sont pas forcément labellisés IA / machine learning / réseaux neuronaux. C'est ainsi le cas du projet **Orpailleur** mené à Nancy et 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 quand à elle directement sur le machine learning et l'auto-apprentissage.

Les chercheurs français **se plaignent** en tout cas d'être délaissés en France dans la discipline. Ils ne sont certainement pas les seuls, au sens où de nombreuses disciplines se sentent délaissées dans la recherche publique.

Une association créée en 1993 fait la promotion de la recherche en IA, l'**AFIA**. Elle organisait en octobre 2014 une **conférence** de promotion de l'IA dans la recherche. On y identifie par exemple **Andreas Herzig** (IRIT, CNRS, Toulouse) qui travaille sur la modélisation de la logique et du raisonnement, **Hélène Fargier** (IRIT, CNRS, Toulouse) qui travaille notamment sur la programmation par contraintes, **Jérôme Euzenat** (LIG, Inria) qui planche sur la représentation et l'échange de connaissances et **Leila Amgoud** (IRIT, CNRS) qui est spécialisée dans la modélisation de l'argumentation.

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** qui est l'un principaux financeurs de projets issus de l'INRIA, on constate que l'IA est souvent en filigrane de ces projets, mais pas forcément au niveau "plateforme" ou "couches de base".



Dans **Economic Report or The President**, le rapport annuel 2016 sur l'économie de la Maison Blanche, j'ai découvert deux données intéressantes : aux USA, 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 milliers. 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 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 et sans objectifs pratiques clairs, sauf lorsqu'elle est financée par des entreprises privées, surtout depuis la loi Pécresse de 2007. A méditer !

Startups horizontales

Voici les startups que j'ai pu repérer dans les solutions techniques d'IA plus ou moins génériques. Le champ de la reconnaissance audio est faiblement couvert par les startups françaises. Dans celle des images, on eu quelques cas anciens comme LTU qui a été racheté par le japonais JASTEC en 2005.

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 paysage 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 réalise maintenant 50% de son chiffre d'affaire.

J’indique comme dans la partie précédente entre parenthèse l’année de création et les montants levés lorsqu’ils sont disponibles. On aimerait bien ajouter un troisième indicateur : le chiffre d’affaire, mais il n’est généralement pas disponible !

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 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.

Dataiku (2013, $3,5m) fait évoluer les concepts de business intelligence et de data mining avec son Data Science Studio, un ensemble d’outils d’analyse de données qui exploitent du machine learning pour la création de modèles de données et de simulations.

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 s’appuie sur force marchine et deep learning. La startup s’appuie sur les travaux de recherche de deux laboratoires publics le CNRS LIP6 and l’ISIR de l’UPMC (Paris VI).

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

Proxem (2007, 1m€) 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.

Smart Me Up (2012, 3m€), vu aux CES 2015 et 2016 propose une solution logicielle d'analyse des visages. Elle détecte l'âge, le comportement et les émotions des utilisateurs. La solution est bien entendu plutôt commercialisée sous forme de brique logicielle en marque blanche utilisable dans des applications métier.

Regaind (2014, 400K€) 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.

Moodstocks (2008) propose une solution mobile de reconnaissance d'images, fournie sous la forme d'APIs et d'un SDK multi-plateforme.

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.

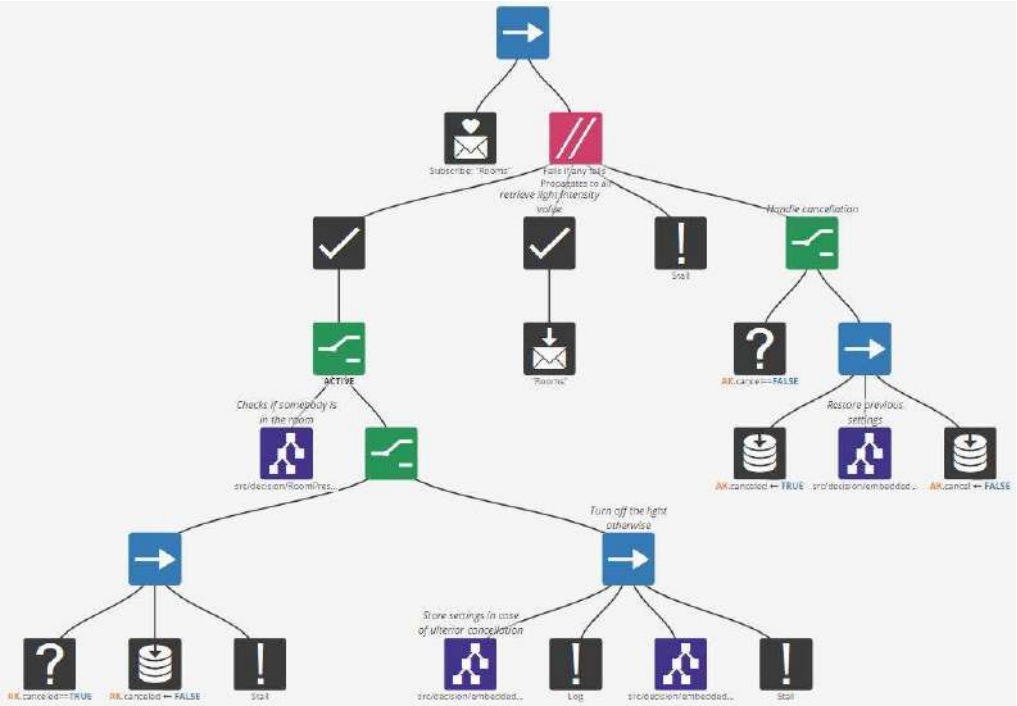
Do You Dream Up (2009) propose également un agent conversationnel multilingues pour les sites web. Il est notamment utilisé par Voyages-SNCF depuis 2011 et a récemment évolué pour être intégré dans une "HelpBox", sorte d'aide en ligne contextuelle interactive.

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 très 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, était également présent au CES de Las Vegas en 2015 et 2016. Il 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 (*ci-dessous*) 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 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 (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. Elle s'appuie sur la reconnaissance vocale, sur la gestion de contexte, sur la détection des intentions et la gestion de règles. Son créateur, Gregory Renard, planche sur l'IA depuis près de 20 ans.

Scortex (2016) développe des solutions matérielles et logicielles apportant l'autonomie aux robots et objets connectés qui intègre notamment la reconnaissance d'images et de la parole. Ils ont même développé un chipset à base de réseaux neuronaux.

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.

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, $3,2m) 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, 1,2m€) 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. Donc, une offre b2b.

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) 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. La startup est accélérée chez Microsoft Ventures à Paris. 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 Technologies (2014) a créé une solution d'interprétation automatique des électrocardiogrammes (ECG) en temps réel s'appuyant sur du machine learning. 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.

Khresterion (2014) propose un logiciel d'aide au diagnostic et à la prescription pour le et les cancers. La solution fonctionne sur un principe voisin de celui d'IBM Watson, compulsant la littérature scientifique et les données des patients pour proposer divers traitements avec leurs avantages et inconvénients comme les effets secondaires. La société aurait comme prescripteur des organismes de remboursement comme Humanis, Axa et la Maaf. Sa solution commence aussi à être utilisée dans la finance, là où les cycles de vente sont probablement plus courts.

Industrie

Il existe probablement de nombreux acteurs dans ce vaste domaine où le machine learning peut avoir plein d'usages.

Citons par exemple le cas de **DCbrain**, issu de Telecom Paristech, et spécialisé dans la modélisation de flux physiques (eau, gaz, électricité, ...) pour les utilities, qui s'appuie sur des solutions de big data et de machine learning pour identifier des tendances et signaux faibles. Cela permet de piloter automatiquement les réseaux, de faire de la maintenance prédictive et de modéliser le fonctionnement des réseaux.

Applications métiers

C'est là que la créativité est la plus développée, comme nous l'avions vu dans l'**article précédent** de la série au sujet des startups américaines.

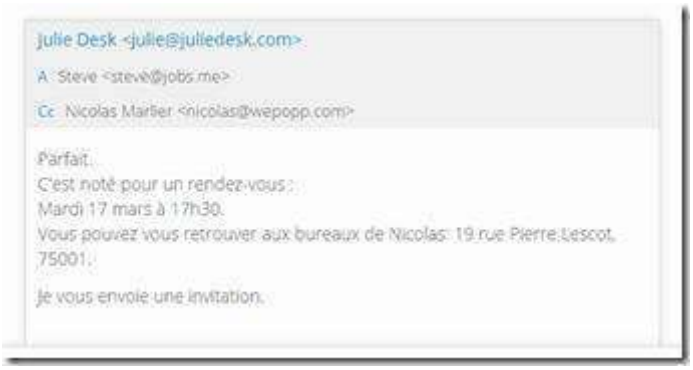
Snips.ai (2013, $6,3m) est une startup connue du secteur de l'IA créée par Rand Hindi (prix du MIT30 en 2015), Mael Primet et Michael Fester. Leur dernière levée de fonds de 5,7m€ en juin 2015 présente la particularité d'associer Bpifrance avec des investisseurs américains, en plus de business angels tels que Brent Hoberman et Xavier Niel. L'équipe comprend 35 personnes : des data-scientists, des développeurs, designers et quelques marketeurs. Leur positionnement est large et un peu vague : rendre la technologie invisible et les usages intuitifs via de l'IA. A ce titre, la startup a développé des applications expérimentales telles que :snips (un ensemble d'applications de recherche pour iOS dont un clavier virtuel intelligent pour la recherche d'adresses), Tranquilien (qui prédit les places disponibles dans les trains de banlieue), Parkr (la même chose pour prédire les places de parking), Flux (qui identifie le trafic mobile en s'appuyant sur les données des smartphones), RiskContext et SafeSignal (identification de risques d'accidents sur la route).

La startup planche aussi sur des applications verticales : pour les véhicules connectés, dans l'hôtellerie, la maison connectée et les loisirs numériques. Le tout s'appuie sur force machine et deep learning, modèles probabilistiques, traitement du langage, gestion de graphes et aussi encryption des données pour garantir la vie privée. Derrière la vision, l'implémentation et l'expérimentation, on leur souhaite de réussir la businessmodelation.

Jam (1m€) a créé un agent conversationnel SMS pour étudiants. Ils ont ISAI Ventures dans leur capital. La solution utilise une combinaison d'IA et de vrais intervenants humains pour assurer une bonne qualité des réponses. Leurs outils d'IA sont en open source.

Julie Desk (2014, $993K), basé à Paris, propose un service d'assistante virtuelle fonctionnant sous la forme d'un agent conversationnel opérant en français et en anglais. Il gère surtout votre agenda et réponds à vos mails à votre place pour prendre des rendez-vous avec vos interlocuteurs. Comme pour Jam, l'agent fonctionne en

mode supervisé par des opérateurs ce qui permet d'assurer une bonne qualité de service. Les tarifs vont de 50€ à 80€ par mois. Il est notamment utilisé par des entrepreneurs de startups. Mais l'agent ne répond pas encore au téléphone.



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 hébergée à l'accélérateur de Microsoft Ventures à Paris 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 autre participation de IT-Translation, la structure de valorisation des projets de recherche issus notamment de l'INRIA. Mais la startup 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.

Yseop (2008) propose son agent conversationnel **Savvy**. Nous l'avions déjà évoqué dans le **troisième article** de cette série. La société propose également une solution de Business Intelligence qui est capable d'extraire intelligemment des données de bases structurées (ERP, CRM, etc) pour les convertir en synthèse en langage naturel, après moulinage dans un moteur de déductions, et qui plus est, dans plusieurs langues.

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 combi-

naison 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!

Dans un **compte-rendu** sur l'écosystème entrepreneurial de La Réunion, j'avais aussi identifié quelques startups qui utilisent le machine learning : **logiCells** (ERP sémantique) et **Teeo** (analyse de consommation d'énergie pour les entreprises). A contrario, certaines startups font appel à des briques d'IA comme le machine learning mais préfèrent ne pas l'évoquer dans leur communication.

Ce tour est probablement incomplet et les oubliés du secteur se feront immanquablement connaitre pour intégrer cette liste que je mettrai à jour au fil de l'eau. A vrai dire, d'ici peu de temps, l'usage de machine learning sera aussi courant 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 !

Modélisation et copie du cerveau

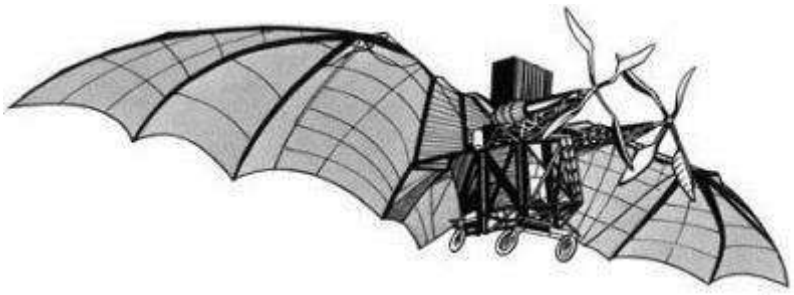
Je vais m'intéresser ici au fonctionnement du cerveau pour en évaluer la complexité et la difficulté à en modéliser le comportement dans de l'IA.

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. 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, 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.

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.

Les initiatives de recherche pour 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.

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 !

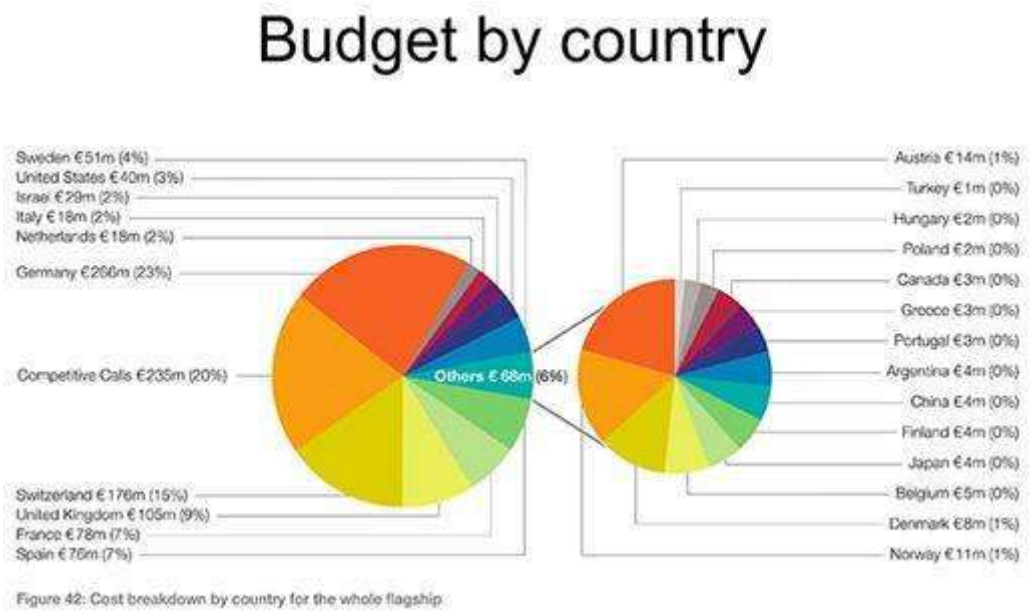
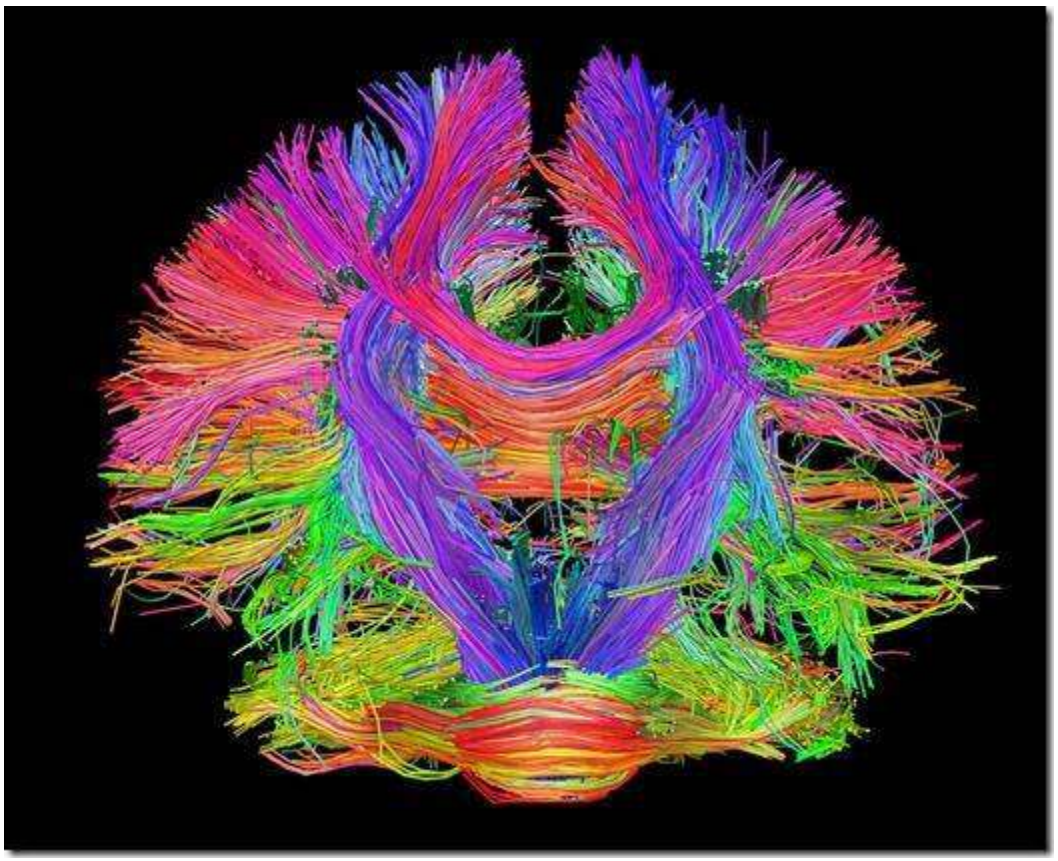


Figure 42: Cost breakdown by country for the whole flagship

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 nano-technologies 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.

La copie du cerveau n'est pas pour demain et heureusement

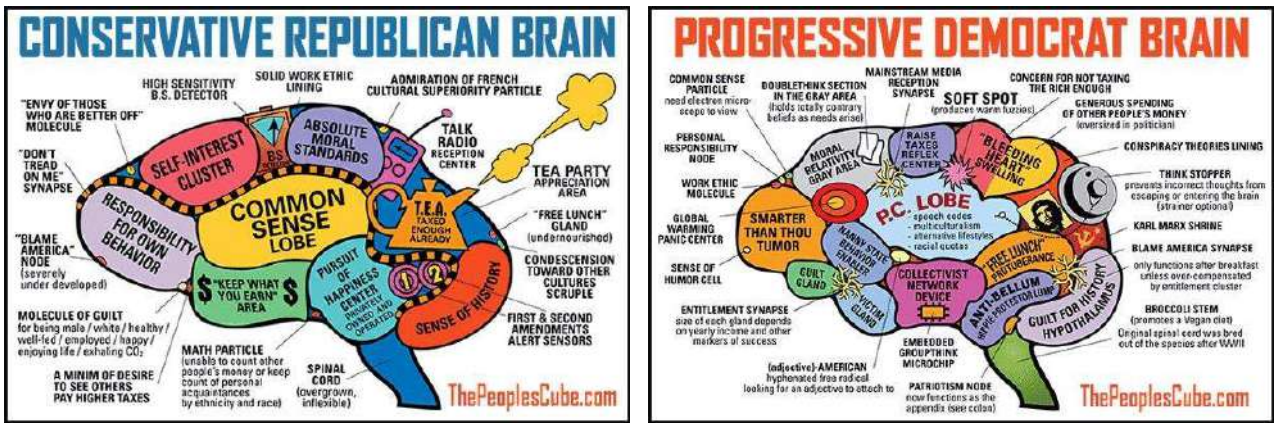
Dans "The Singularity is Near", Ray Kurweil fantasme sur la capacité à venir de transplanter un cerveau dans une machine et d'atteindre ainsi l'immortalité, incarna-

tion 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 !

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[18].

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 !



La science progresse aussi pour cartographier le cerveau politique des électeurs. Ici, sur le clivage républicains et démocrates aux USA qui est toujours d'actualité. Ce n'est évidemment pas sérieux, mais un jour, cela le sera peut-être et on pourra cartographier le cerveau des électeurs ! **Source**.

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

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

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 ?

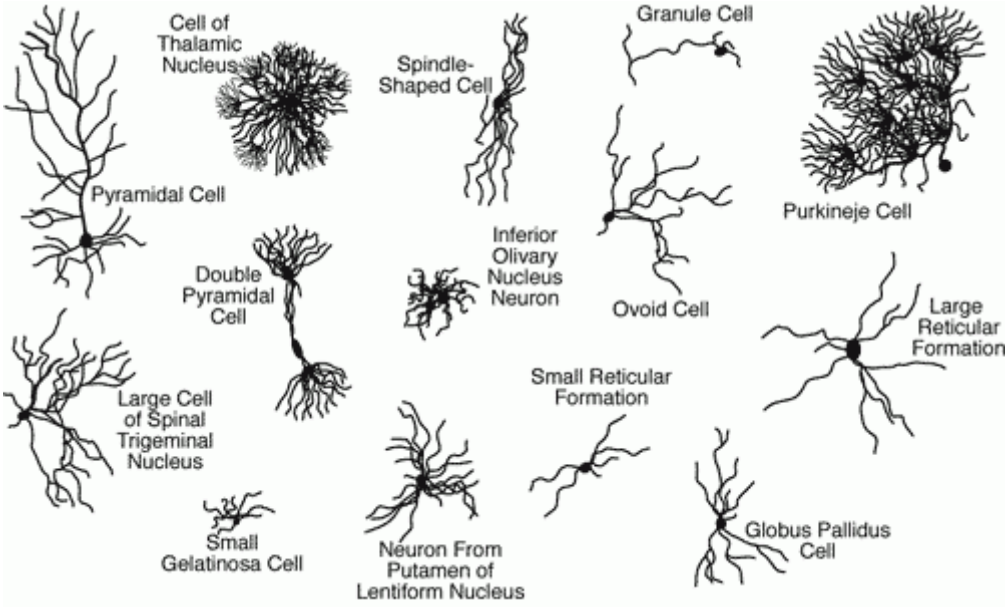
Il n'y a pas 36 solutions : il faut passer par des ondes électromagnétiques, et avec une précision de l'échelle du nanomètre. Aujourd'hui les scanners utilisent généralement trois technologiques : la tomodensitographie qui mesure la densité de la matière par rayons X, les PET scanners qui détectent des traceurs biologiques radioactifs par émission de photons et l'IRM qui détecte les corps mous par résonnance magnétique 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 ! Il serait d'ailleurs intéressant d'évaluer la quantité d'énergie qu'il faudrait envoyer dans le cerveau (rayons X, magnétisme, etc) pour réaliser ce genre de détection 100 à 1000 trillions de fois, soit le nombre de synapses dans le cerveau, et avec une résolution de l'ordre du nanomètre. Si cela se trouve, cette énergie électromagnétique serait suffisante pour faire cuire le cerveau comme dans un four à micro-ondes, ce qui ne serait pas du tout un effet désiré ! Sauf peut-être pour traiter de manière un peu radicale certains abrutis.

C'est une loi connue dans la physique : plus on explore l'infiniment petit, plus c'est couteux en énergie. Le LHC du CERN près de Genève a permis de détecter les bosons de Higgs. Il consomme la bagatelle de **200 mégawatts** avec des pointes de 1,3 gigawatts, soit plus que la puissance générée par une tranche de centrale nucléaire ! Le LHC a couté $9B et le prix de ce genre d'instrument scientifique ne suit pas du tout la loi de Moore comme le séquençage de l'ADN !

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[19].

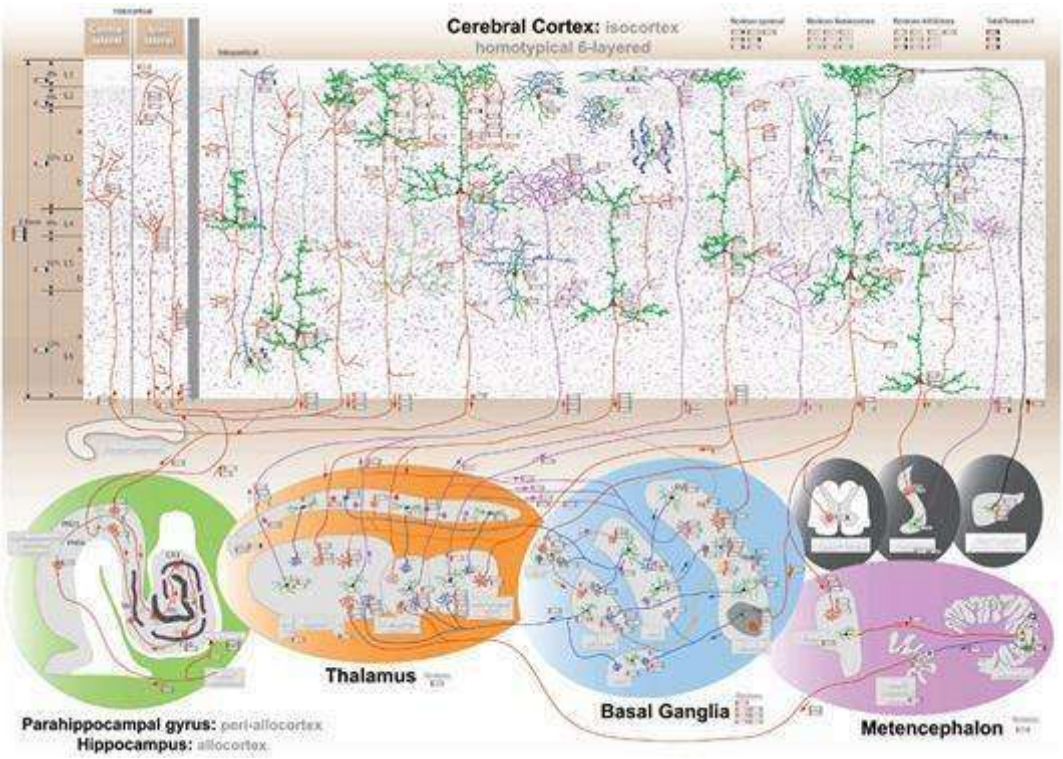


(source du schéma ci-dessus sur quelques exemples de neurones cérébrales : **http://neuromorpho.org**)

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.

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 !

[19] 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.

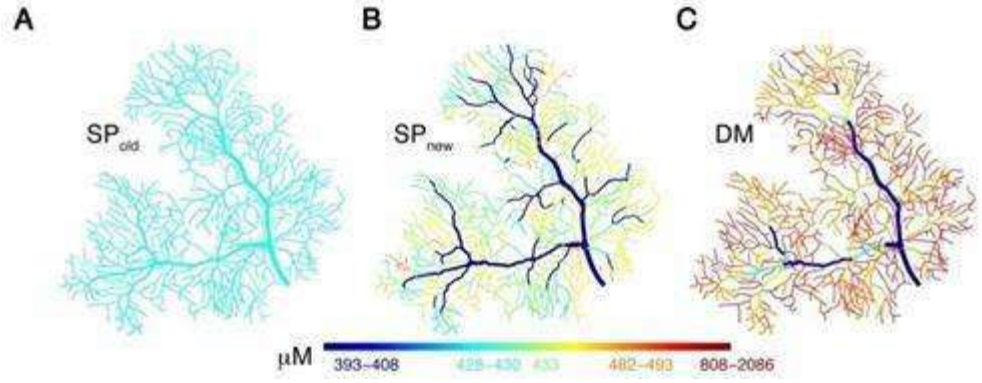


(source de l'illustration ci-dessus)

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

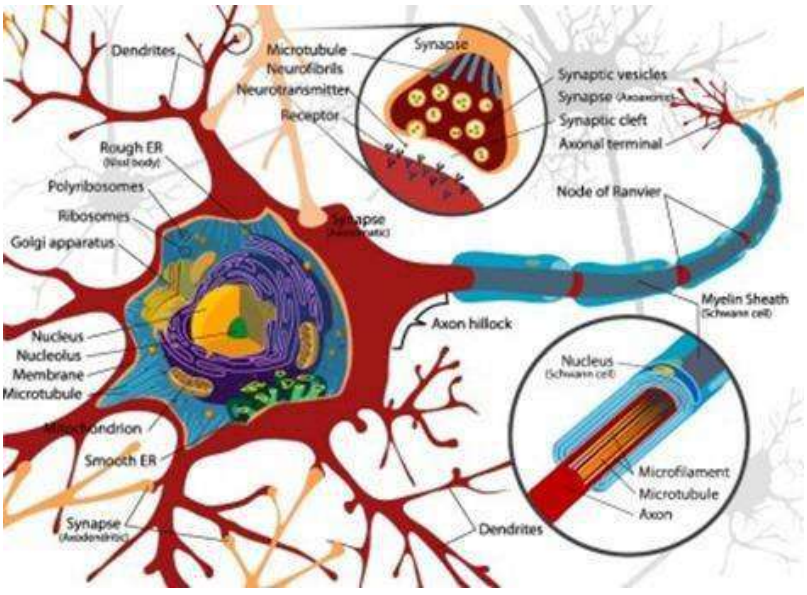
Un cerveau adulte comprend environ 85 milliards de neurones dont 30 dans le cortex, 10 trillions de synapses (liaisons neurones / neurones via les terminaisons multiples des axones qui sortent de neurones et se connectent aux dendrites proches du noyau d'autres neurones), et 300 milliards de dendrites (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.

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.

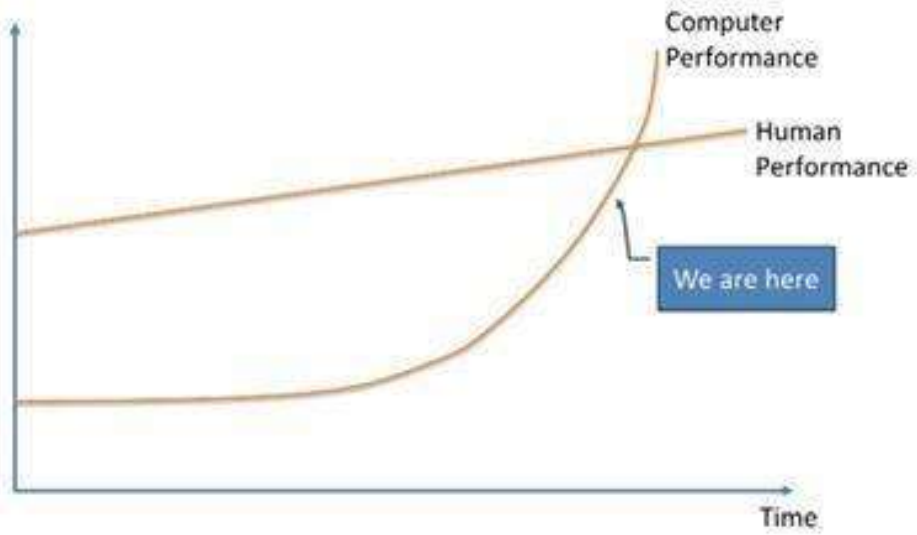


(**source** du schéma qui l'explique très bien)

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.

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 (*ci-dessus*).

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

Evolutions de la loi de Moore et applications à l'intelligence artificielle

L'IA a connu des vagues diverses d'hivers et de renaissances. Pour certains, il s'agit plutôt de vaguelettes. Les récentes "victoires" de l'IA comme dans Jeopardy (2011) et AlphaGo (2016) donnent l'impression que des sauts quantiques ont été franchis. C'est en partie une vue de l'esprit car ces progrès sont sommes toutes modestes et réalisés dans des domaines très spécialisés, surtout pour le jeu de Go.

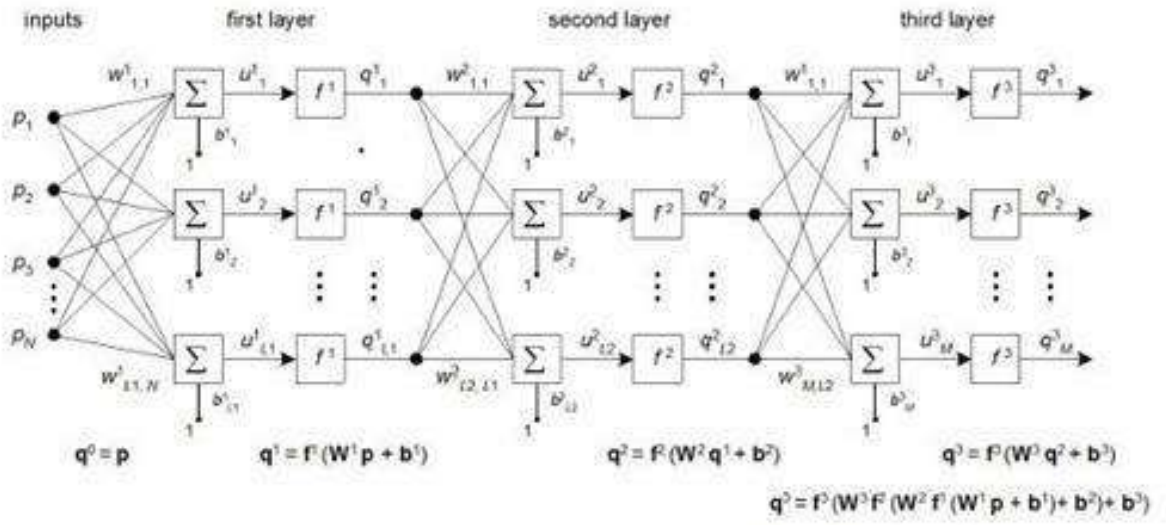
Peut-on décortiquer par quels biais les progrès dans l'IA vont s'accélérer ? Nous avons vu dans les parties précédentes qu'il était difficile de faire la part des choses entre avancées liées à l'immatériel et celles qui dépendent du matériel. Je vais commencer par les algorithmes et logiciels puis aborder la partie matérielle. Avec en interlude, un passage sur l'application de la loi de Moore dans la vraie vie qui est bien différente des belles exponentielles présentées à tout va !

Algorithmes et logiciels

Nous verrons plus loin que le matériel continuera de progresser, même si c'est un chemin semé d'embuches du côté des processeurs.

S'il y a bien une loi de Moore difficile à évaluer, c'est celle des algorithmes et logiciels ! Personne ne la mesure et pourtant, une bonne part des progrès numériques vient de là et pas seulement de l'augmentation de la puissance du matériel.

Les réseaux neuronaux à boucle de feedback et le deep learning auto-apprenants sont maintenant anciens et leur progression est lente dans le principe. Leur mise en œuvre s'améliore beaucoup grâce aux possibilités matérielles qui permettent de créer des réseaux neuronaux multicouches allant jusqu'à 14 couches.



A chaque fois qu'un record est battu comme avec AlphaGo, il résulte de la combinaison de la force du matériel, du stockage et du logiciel. Qui plus est, ces records de

l'IA portent sur des domaines très spécialisées. La variété et les subtilités des raisonnements humains sont encore loin. Mais elles ne sont pas hors de portée. Notre cerveau est une machine hyper-complexe, mais ce n'est qu'une machine donc potentiellement imitable.

La recherche progresse en parallèle dans les techniques de reconnaissance d'images (à base de réseaux de neurones et de machine learning), de la parole (itou) et de l'analyse de données (idem). Les algorithmes génétiques sont de leur côté utilisés pour trouver des chemins optimums vers des solutions à des problèmes complexes intégrant de nombreux paramètres, comme pour trouver le chemin optimum du voyageur du commerce.

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 un 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. Un nouveau savoir est à créer : le benchmark de solutions d'IA ! Voilà un métier du futur !

La **vie artificielle** est un autre pan 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.

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 con-

naissances – 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, on en est encore loin. Qui plus est, ce n'est pas forcément nécessaire pour résoudre des problèmes courants de l'univers des entreprises. Comme l'indique si bien **Yuval Noah Harari**, l'auteur du best-seller "Sapiens"[20], "*L'économie a besoin d'intelligence, pas de conscience*" ! Laissons donc une partie de notre intelligence voire une intelligence plus développée aux machines et conservons la conscience, les émotions et la créativité !

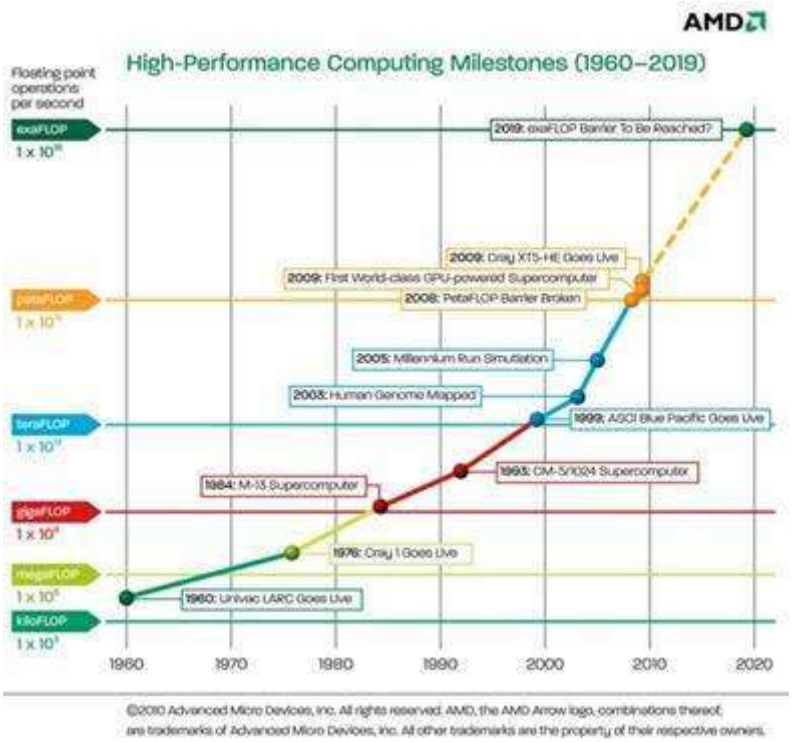
La loi de Moore dans la vraie vie

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.

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.

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



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 les mêmes.

Examinons donc de près comment cette fameuse loi s’applique pour des objets numériques grand public. Prenons trois cas d’usages courants : un **laptop** plutôt haut de gamme en 2006 et en 2016, l’évolution de l’**iPhone** entre sa première édition lancée en juin 2007 et l’iPhone 6S lancé en septembre 2015 et puis l’évolution du **haut débit fixe** sur 10 ans.

En appliquant une belle loi de Moore uniforme, les caractéristiques techniques de ces trois larrons devraient doubler tous les deux ans au minimum. Sur une période de 10 ans, cela donnerait 2 puissance 5 soient **x32** et sur 8 ans, **x16**. Si le doublement intervenait tous les 18 mois, ces facteurs seraient respectivement de **x101** et **x40**.

Commençons par un **laptop** haut de gamme à prix équivalent entre 2006 et 2016. J’ai comparé deux modèles plutôt haut de gamme de la même marque : un Asus W7J de 2006 et un Asus Zenbook UX303UA de 2016, certes sorti en 2015. Mais entre fin 2015 et mi 2016, il n’y a pas eu de changements d’architecture des laptops, qui collent à la roadmap d’Intel.

Aucun paramètre technique n’a évolué d’un facteur x32 et à fortiori d’un facteur x100. Ceux qui ont le mieux progressé et qui ont un impact sur la performance perçue par l’utilisateur sont la vitesse du moteur graphique (x12) et celle du Wi-Fi (x24).

Pour le reste, les gains sont très modestes. Le processeur est “seulement” 3,7 fois plus rapide. La résolution des écrans a augmenté mais la résolution limitée de l’œil rend caduque cette progression dès lors qu’un écran atteint la résolution 4K, qui commence à apparaitre sur certains laptops.

Le plus grand retardataire est la batterie qui n’évolue quasiment pas. L’autonomie des laptops a progressé non pas grâce aux batteries mais à la baisse de consommation des processeurs et autres composants électroniques ainsi qu’à l’intelligence intégrée dans les systèmes d’exploitation, aussi bien Windows que MacOS.

Les derniers processeurs Intel savent éteindre certaines de leurs parties lorsqu’elles ne sont pas utilisées. Par contre, la densité des batteries s’est un peu améliorée et leur cure d’amaigrissement a permis de créer des laptops plus fins.

	2006	2016	Multiple de performance	Commentaire
PC laptops servant à la comparaison	**Asus W7J**	**Asus Zenbook UX303UA**		Deux laptops plutôt haut de gamme au moment de l'évaluation.
CPU	Intel Core 2 Duo T7200	Intel Core i7 6500		Processeurs mobiles haut de gamme de laptop pour ces deux périodes.
Cœurs	2	2	1	Stabilité du nombre de cœurs. Elle augmente plutôt sur les CPU de desktops.
Clock	2 GHz	2,5 GHz	1,25	Faible évolution de l'évolution de la vitesse d'horloge.
Gravure et évolution de la densité	65 nm	14 nm	21,6	Le nombre de transistors augmente, notamment du fait du GPU.
Transistors	291 millions	1,5 milliards (estimation)	5,2	Intel ne communique plus sur le nombre de transistors par chipset depuis 2015.
Cache	4 Mo	4 Mo	1	Rien de changé.
Performance en Passmark	1175	4328	3,7	Mesure la plus importante de performance, perçue par l'utilisateur.
Consommation (TDP)	34W	7.5 à 25 W	4,5	Baisse de la consommation qui permet d'augmenter l'autonomie.
Mémoire RAM	1,5 Go	8 Go	5,3	Contribue à améliorer la vitesse de passage d'une application à l'autre.
Fréquence mémoire RAM	533 MHz	1600 MHz	3	Permet de transférer plus rapidement les données en mémoire.
GPU	GeForce 7400 128 Mo	Intel HD Graphics 520		Processeur graphique.
G3D Benchmark	63	768	12,2	Belle évolution sachant que le GPU intégré aux chipsets Intel n'est pas ce qu'il y a de mieux pour un laptop vs une carte nVidia ou AMD.
Stockage	120 Go HDD 5400 rpm	512 Go SSD	4,2	Augmentation modérée du fait du passage du HDD au SSD. Avec un HDD de 2 To, le ratio serait de 17.
Bus stockage	SATA 1 (150)	SATA 3 (600)	4	6 Gbits/s sachant qu'avec un port M.2, on peut aller jusqu'à 1,6 Go/s.
Ethernet	1 Gbits/s	1 Gbits/s	1	Rien de changé.
Wi-Fi	802.11 g à 54 Mbits/s	802.11 ac à 1300 Mbits/s	24	La plus grande évolution dans un laptop !
Résolution écran	1 366×768	3200x1800	5,49	Belle évolution mais qui sera limitée à terme par la résolution oculaire.
Résolution webcam	1,3 Mpixels	720p		Pas d'évolution.
Batterie	49 Wh	56 Wh	1,14	Les batteries évoluent peu. La consommation baisse, mais pas l'énergie disponible.
Poids	1,95 Kg	1,3 Kg	1,50	Ca s'allège doucement.
Epaisseur	36,9 mm	19,2 mm	1,92	Ca maigrit bien. L'impact du Macbook Air lancé en 2010.
Prix	1 800 €	1 560 €	1,15	Prix relativement stable.

Du côté de l’**iPhone**, la situation est plus contrastée et bien meilleure que pour les laptops.

Deux dimensions techniques ont bien progressé : le processeur qui est x18 fois plus rapide et la communication data Internet mobile qui est **x781** fois plus rapide, tout du moins en théorie, car d'un point de vue pratique, le ratio réel est plus raisonnable. Et encore, c'est lié au choix peu hardi du Edge au moment de la sortie du premier iPhone alors que la 3G aurait très bien pu être supportée, réduisant ainsi l'écart de performance entre l'iPhone 1 et le 6S.

Contrairement aux laptops, au lieu de voir les prix baisser, ils augmentent, positionnement haut de gamme d'Apple oblige. Le poids augmente aussi car l'iPhone 6S a un écran plus grand que celui du premier iPhone. Et comme pour les laptops, la capacité de la batterie a très peu augmenté. J'ai indiqué les résolutions d'écrans et de capteurs vidéo sachant qu'il n'y a pas de raison objective de vouloir poursuivre ad-vitam la loi de Moore pour ce qui les concerne.

	Juin 2007	Septembre 2015	Multiple de performance	Commentaire
	iPhone 1	iPhone 6S		
CPU	Samsung S5L8900	Apple A9		Les processeurs des iPhone sont conçus par Apple et fabriqués par Samsung et TSMC.
Cœurs et clock	un cœur 32 bits à 412 MHz	double cœur, 64 bits 1,8 GHz	18	Meilleure évolution de la clock grâce aux noyaux ARM basse consommation.
Gravure	90 nm	14 nm	41,3	
GPU	PowerVR MBX Lite 60 MHz	PowerVR GT7600		
Stockage max	16 Go	128 Go	8	Progression des mémoires Flash continue.
Wifi	Wi-fi g à 54 Mbits/s	Wi-fi ac à 866 Mbits/s	16	Progression moindre que sur PC car le ac s'applique à géométrie variable.
WAN	Edge à 384 kbit/s	LTE à 300 Mbits/s	781	La plus forte évolution technologique inventoriée ici !
Ecran	480 x 320	1334*750	6,5	Meilleure progression avec le 6 Plus et aussi Sony Xperia Z5 en 4K.
Capteur dorsal	2 Mpixels	12 Mpixels	6	Certains smartphones atteignent 20 Mpixels mais cela ne sert à rien.
Batterie	1400 mAh	1642 mAh	1,2	Très faible évolution. Le 6S Plus a une batterie de 2900 mAh.
Poids	135 g	143 g	0,94	Le 6S est plus lourd car l'écran est plus grand.
Prix	659 € pour 16 Go	1079 € pour 128 Go 859 € pour 16 Go	0,6	Régression, eu égard à la politique de prix et de marge d'Apple.

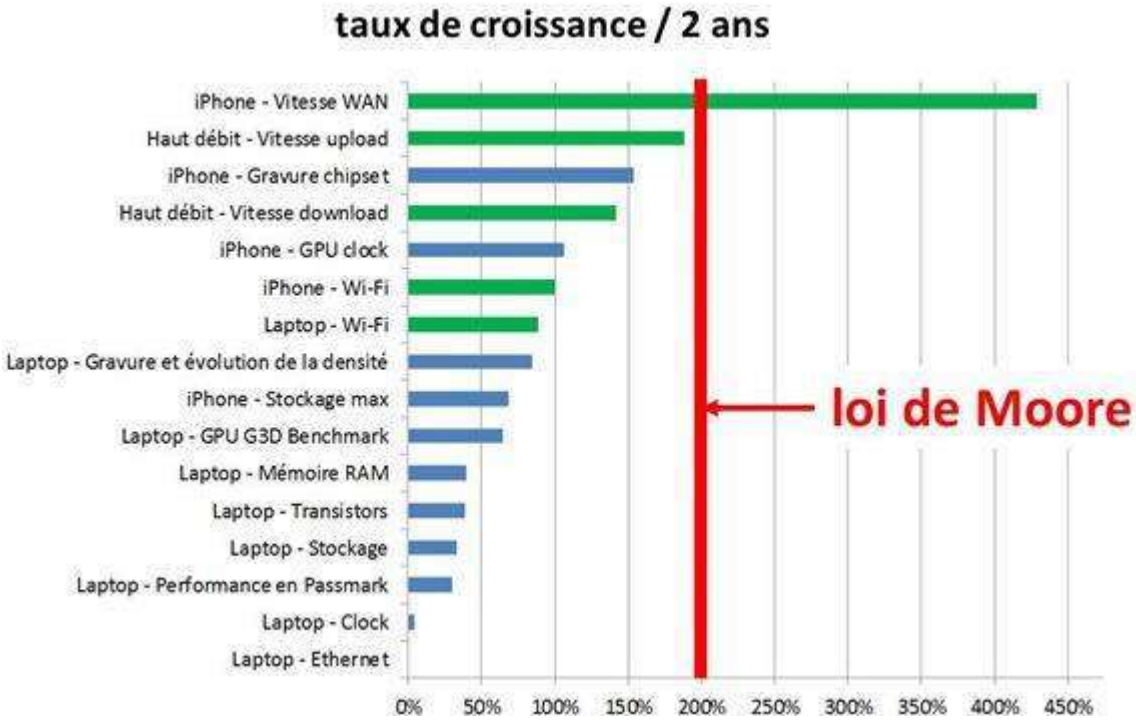
La situation est assez différente du côté du **haut débit fixe**. Vous pouvez stagner pendant une décennie à la même vitesse d'accès à Internet et bénéficier tout d'un coup d'un progrès soudain appliquant 10 ans de loi de Moore. Si vous passez par exemple d'un ADSL à 12 Mbits/s en download et 1 Mbits/s en upload à de la fibre

chez Free à 1 Gbits/s en download et 200 Mbits/s en upload, le facteur multiplicateur est respectivement de x83 et x200.

Si vous partiez d'un débit encore plus faible du fait d'un plus grand éloignement des centraux télécoms, le facteur multiplicateur serait encore plus élevé. Mais plus votre débit ADSL d'origine est faible, plus faibles sont les chances de voir la fibre arriver chez vous du fait des travaux d'infrastructure à réaliser pour passer les fourreaux transportant la fibre du central télécom jusqu'à chez vous !

Chez les autres opérateurs que Free, le facteur multiplicateur dépend de la technologie utilisée. Chez Numericable, c'est du FTTB à la performance à géométrie variable selon l'âge du capitaine et surtout un débit montant assez limité. Chez Orange, vous avez des taquets de débits à 100, 200 et 500 Mbits/s en download et de 50 Mbits/s à 200 MBits/s en upload selon l'offre commerciale. Et si vous attendez toujours la fibre, la loi de Moore vous concernant est un encéphalogramme tout plat !

En ne conservant que les paramètres techniques où la loi de Moore est pertinente, voici donc ce que cela donne sous une autre forme, à savoir la progression moyenne tous les deux ans. On voit qu'à part la data WAN, on est loin du doublement tous les deux ans de la performance !



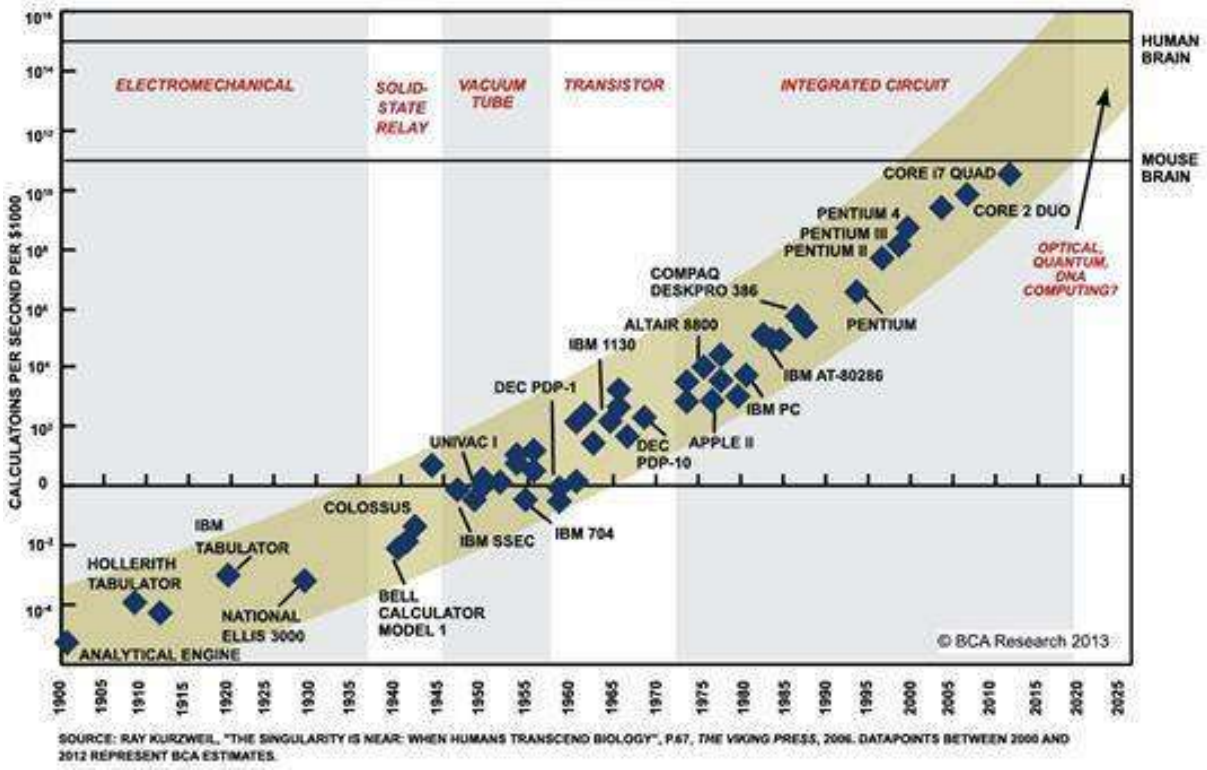
La loi de Moore s'applique bien mieux aux liaisons réseaux haut débit fixe et mobiles qu'à la capacité de calcul et de stockage, surtout sur ordinateurs personnels. Cela explique indirectement la montée en puissance des architectures en cloud. On peut plus facilement répartir une grosse charge de calcul sur des serveurs que sur des postes de travail ou des mobiles. On retrouve cette architecture dans Siri qui traite une bonne part de la reconnaissance vocale côté serveurs. Au passage, la loi de Moore de la vraie vie valide aussi le scénario de fiction de "Skynet" des films Terminator où c'est

une intelligence logicielle distribuée sur des millions de machines dans le réseau qui provoque une guerre nucléaire !

Alors, la loi de Moore est foutue ? Pas si vite ! 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. Mais son redémarrage pourrait prendre du temps.

Puissance de calcul

La fameuse loi de Moore est mise en avant par les singularistes pour prédire le dépassement de l'homme par l'IA à une échéance de quelques décennies. Seulement voilà, la validation dans la durée de cette loi empirique de Moore n'est pas triviale comme nous venons de le voir.



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 est déjà à 10 nm à ce jour, notamment chez Intel, et à 14 nm en version commerciale (Core M et Core i de génération Skylake 2015). 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.

Le schéma ci-dessus et qui vient de Ray Kurzweil n'a pas été mis à jour depuis 2006. Il est difficile d'obtenir un schéma sur l'application de la loi de Moore au-delà de 2010 pour les processeurs. Est-ce parce que l'évolution de la puissance de calcul s'est calmée depuis ?

Dans le même temps, les découvertes en neuro-biologies évoquées précédemment augmentent de plusieurs ordres de grandeur la complexité de la modélisation du fonctionnement d'un cerveau humain. Bref, cela retarde quelque peu l'échéance de la singularité.

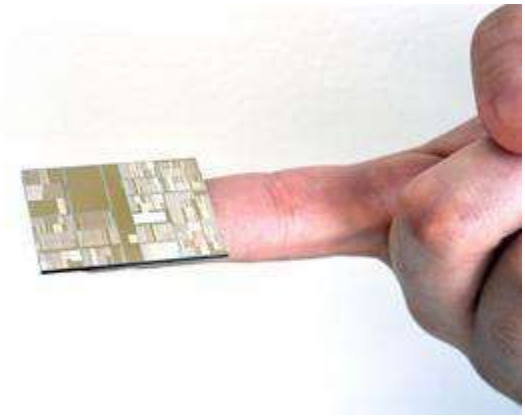
L'excellent dossier **After Moore's Law**, paru dans The Economist en mars 2016, détaille bien la question en expliquant pourquoi la loi de Moore des transistors CMOS pourrait s'arrêter d'ici une douzaine d'année lorsque l'on descendra au niveau des 5 nm d'intégration. Et encore, la messe n'est pas encore 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. Mais l'ingénuité humaine a des ressources insoupçonnables comme elle l'a démontré dans les générations précédentes de processeurs CMOS !

Il faudra tout de même trouver autre chose, et en suivant divers chemins de traverse différents des processeurs en technologie CMOS.

Voici les principales pistes connues à ce jour et qui relèvent toutes plutôt du long terme :

Continuer à descendre coute que coute le niveau d'intégration

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 à ce jour descendu à 10 nm. 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.

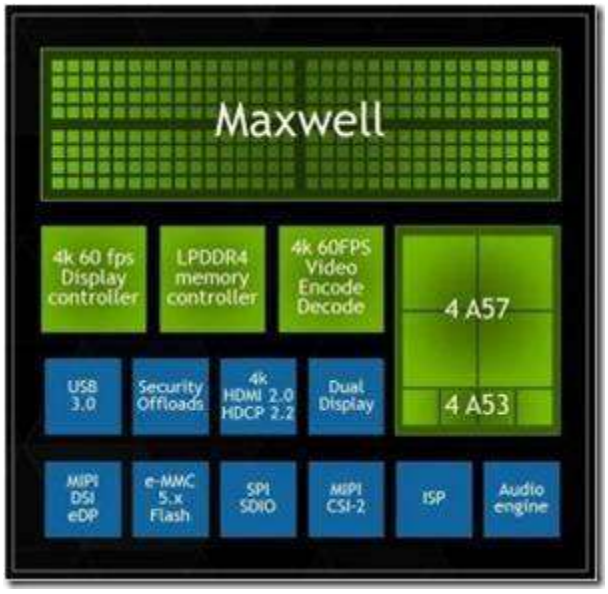


Le multi-patterning, que j'explique **ici**, permet d'en contourner les limitations. Mais il coute tout aussi cher car il ajoute de nombreuses étapes à la fabrication des chipsets et peut augmenter le taux de rebus. La loi de Moore s'exprime en densité de transistors et aussi en prix par transistors.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

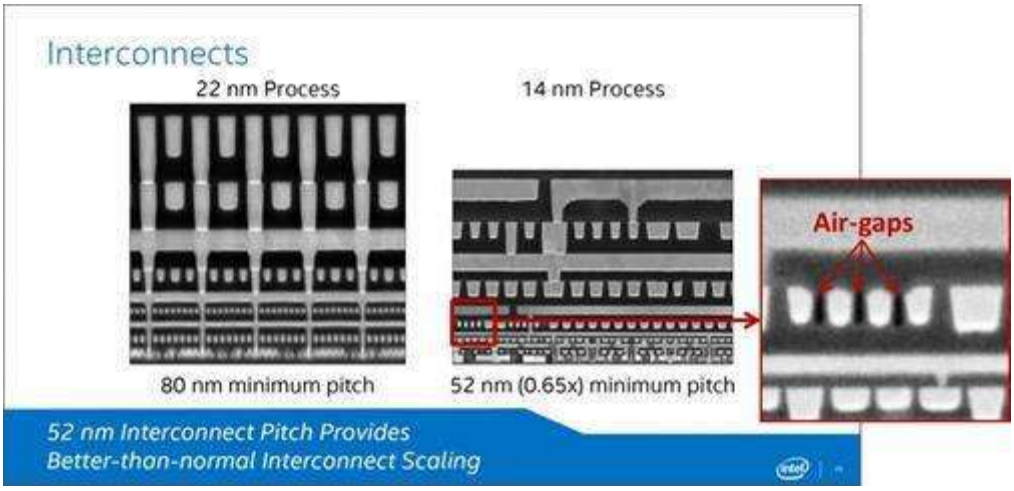
Créer des processeurs spécialisés

Ils sont notamment utiles pour créer des réseaux neuronaux, comme nous l'avions déjà vu dans la **seconde partie** de cette série. La piste est intéressante et est déjà très largement utilisée dans le cadre des GPU ou des codecs vidéo qui sont souvent décodés dans le matériel et pas par logiciel, comme le format HEVC qui est utilisé dans la diffusion de vidéo en Ultra Haute Définition (4K).

C'est l'approche de Nvidia avec ses chipsets X1 (*ci-dessous*) à 256 cœurs ou plus, qui sont utilisés dans la reconnaissance d'images des véhicules autonomes ou à conduite assistée comme les Tesla S. Ces GPU simulent des réseaux neuronaux avec une faculté d'auto-apprentissage. La piste se heurte cependant aux limites de la connectique. Pour l'instant, dans les réseaux neuronaux matériels, chaque neurone n'est relié qu'à ceux qui sont avoisinants dans leur plan. Dans le cerveau, l'intégration des neurones est tridimensionnelle.



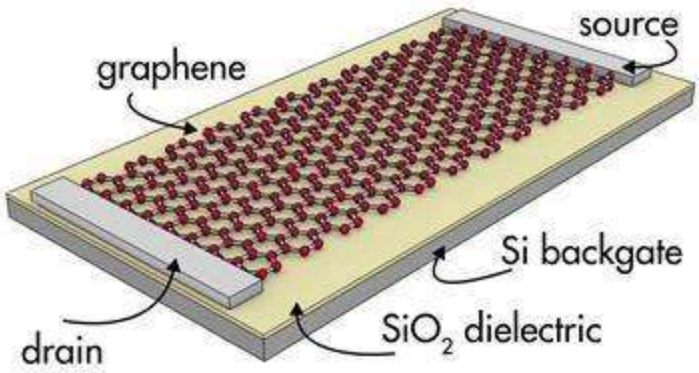
Il est possible d'imiter cette architecture 3D avec des couches métalliques multiples dans les circuits intégrés mais elles coutent pour l'instant assez cher à produire et plus on les empile, plus cela devient compliqué. Les processeurs les plus modernes comprennent une petite dizaine de couches de métallisation, comme indiqué dans ce schéma d'origine Intel.



Il n'est cependant pas théoriquement impossible de superposer des processeurs les uns sur les autres, tout du moins, tant que l'on peut limiter leur réchauffement. L'empilement serait concevable en baissant la fréquence des chipsets, ou avec des techniques extrêmes de refroidissement. Même en divisant par mille la clock des chipsets CMOS, ils resteraient bien plus rapides que la "clock" du cerveau qui est de l'ordre de 100 Hz.

Changer de technologie au niveau des 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.



Passer de l'électron au photon

C'est la photonique qui exploite des composants à base des matériaux dits "III-V"[21]. 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** 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.

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.

Plancher sur les 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 !

Mais sauf à être un spécialiste du secteur, on n'y comprend plus rien ! 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. Vous êtes très fort si vous comprenez quelque chose à partir de la fin de la seconde page ! Et la presse généraliste et même scientifique simplifie tellement le propos que l'on croit avoir compris alors que l'on n'a rien compris du tout !

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 quan-

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

tiques 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.

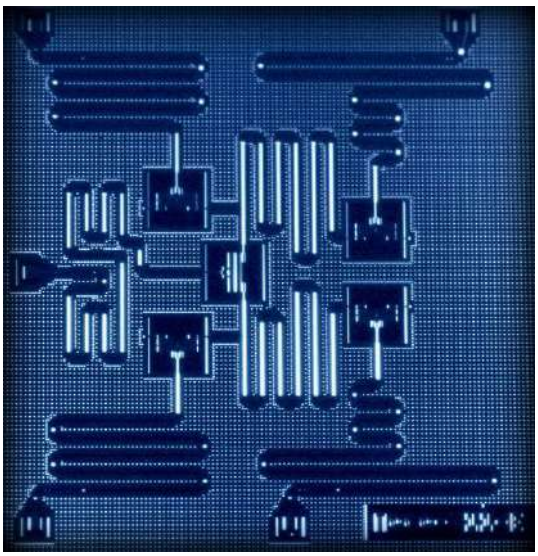
En 2014, des chinois de l'Université de Sciences et Technologies de Hefei ont été parmi les premiers à **expérimenter** des ordinateurs quantiques pour mettre en jeu des réseaux de neurones artificiels, pour la reconnaissance d'écriture manuscrite. Leur ordinateur quantique utilise un composé organique liquide associant carbone et fluor. On n'en sait pas beaucoup plus !

Les équipes de la NASA ont créé de leur côté le QuAIL, le **Quantum Artificial Intelligence Laboratory**, en partenariat avec Google Research. Il utilise un D-Wave Two comme outil d'expérimentation, à ce jour le seul ordinateur quantique commercial, diffusé à quelques unités seulement. 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. Il se pourrait même que ces algorithmes soient codés "en dur" dans les qubits des D-Wave ! 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.

Bref, 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.

Début mai 2016, **IBM** annonçait mettre à disposition son ordinateur quantique expérimental cryogénique de 5 Qubits en ligne dans son offre de cloud. On ne sait pas trop quel type de recherche pourra être menée avec ce genre d'ordinateur ni quelles APIs sont utilisées.



Quid des recherches en France ? 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. Ces composants ont toutefois besoin d'être refroidis près du zéro absolu (-273°C) pour fonctionner. 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**.

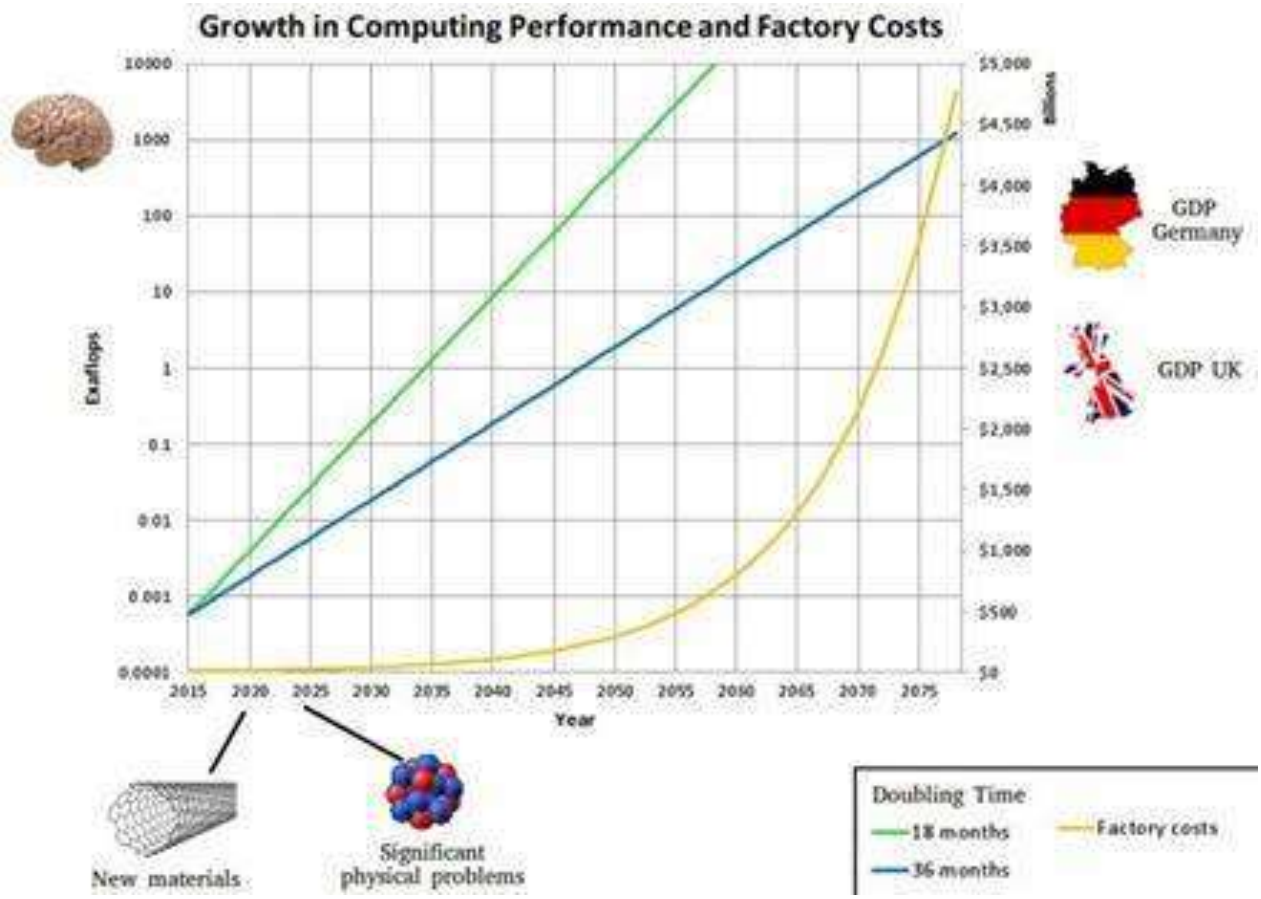
Dans son étude **Quantum Computing Market Forecast 2017-2022**, le cabinet Market Research Media prévoit que le marché des ordinateurs quantiques fera $5B d'ici 2020, en intégrant toute la chaine de valeur matérielle et logicielle. Le premier marché serait celui de la cryptographie. Avant de parler de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

Explorer les 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 (**source**). 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 en bon et vieux CMOS !

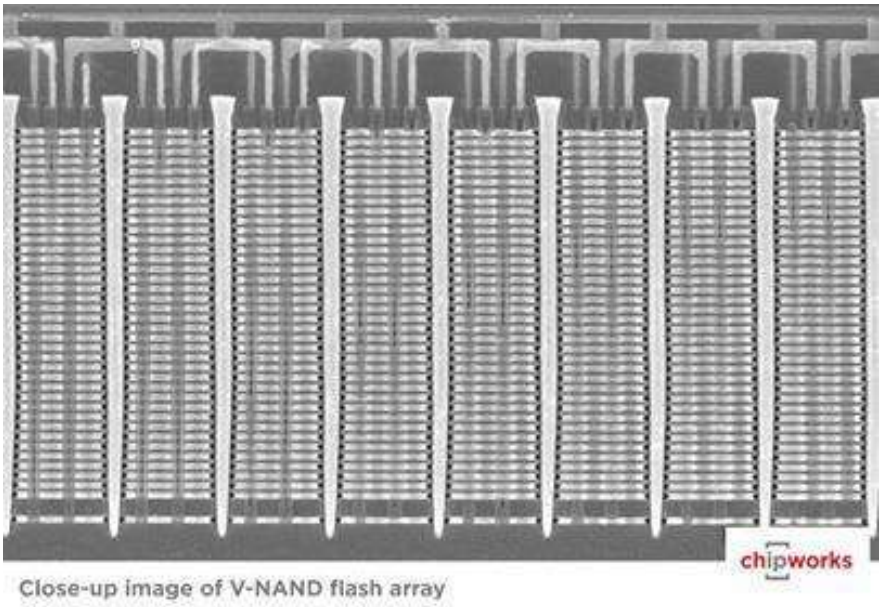
Ceci est d'autant plus plausible que certaines techniques sont insuffisantes pour créer un ordinateur générique, notamment les ordinateurs quantiques qui ne sauraient gérer qu'une certaine classe de problèmes, mais pas comprimer ou décomprimer une vidéo par exemple, ou faire tourner une base de données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

Stockage

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 8 To. Donc, 2 puissance 4 et Moore est sauf. L'évolution s'est ensuite déplacée vers les disques SSD à mémoires NAND dont la capacité, démarrée plus bas que celle des disques durs, 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.

Comme nous l'avions survolé dans le dernier **Rapport du CES 2016**, les mémoires NAND 3D font des progrès énormes, notamment avec 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. Elle est encore à l'état de prototype mais sa fabrication ne semble pas hors de portée.

La technologie de mémoire 3D est aussi maîtrisée par des sociétés telles que **Samsung** (ci-dessous, avec sa technologique V-NAND) et **Toshiba** (ci-dessus avec sa technologie BiCS). Elle consiste à créer des puces avec plusieurs couches empilées de transistors, ou de transistors montés en colonnes. L'e niveau d'intégration le plus bas des transistors est ici équivalent à celui des CPU les plus denses : il descend jusqu'à 10 nm.



Close-up image of V-NAND flash array

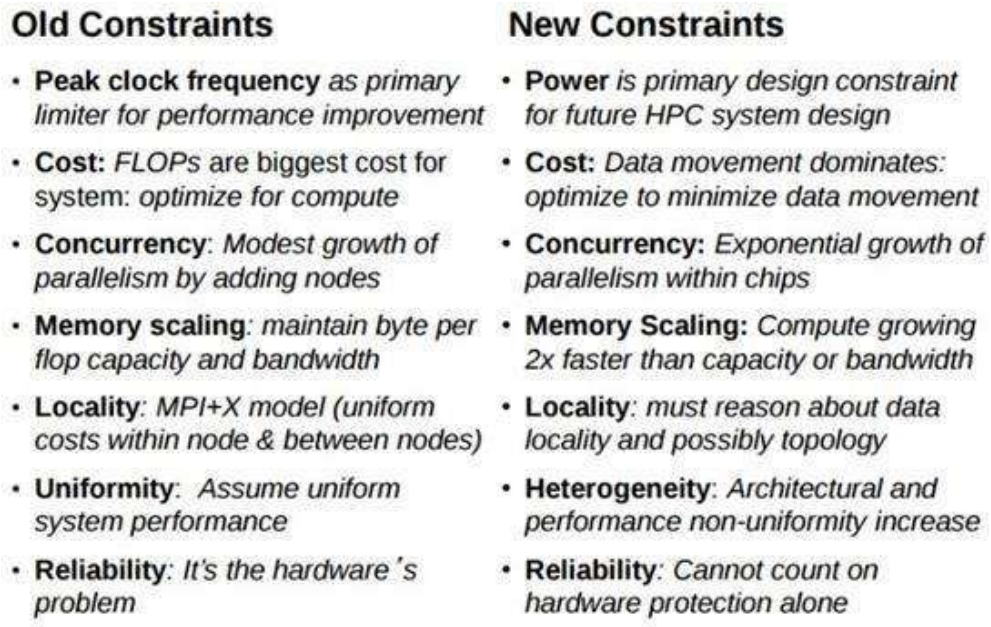
On sait empiler aujourd'hui jusqu'à 48 couches de transistors, et cela pourrait rapidement atteindre une centaine de couches.

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.

Pour les supercalculateurs, une tâche ardue est à accomplir : accélérer la vitesse de transfert 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 1,6 Go/s, soit un dixième de ce qui est recherché dans les calculateurs à haute performance (HPC). Mais cette vitesse semble supérieure à celle de lecture d'un SSD ! Le bus de communication est devenu plus rapide que le stockage !

Avec 3D XPoint, l'accès aux données serait 1000 fois plus rapide qu'avec les SSD actuels, modulo l'interface utilisée. Après un retard à l'allumage, cette technologie pourrait voir le jour commercialement en 2017. Elle aura 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 !

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.



(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[22]. La densité d'un tel stockage serait énorme. Son avantage est sa durabilité, estimée à des dizaines de milliers d'années, voire plus selon les techniques de préservation. Reste à trouver le moyen d'écrire et de lire dans de l'ADN à une vitesse raisonnable.

Aujourd'hui, on sait imprimer des bases d'ADN à une vitesse incommensurablement lente par rapport aux besoins des ordinateurs. Cela se chiffre en centaines de bases par heure au grand maximum. Cette vitesse s'accélèrera sans doute dans les années à venir. Mais, comme c'est de la chimie, elle sera probablement plus lente que les changements de phase ou de magnétisme qui ont court dans les systèmes de stockage numérique actuels. La loi de Moore patientera donc quelques décennies de ce côté là, tout du moins pour ses applications dans le cadre de l'IA.

Capteurs sensoriels

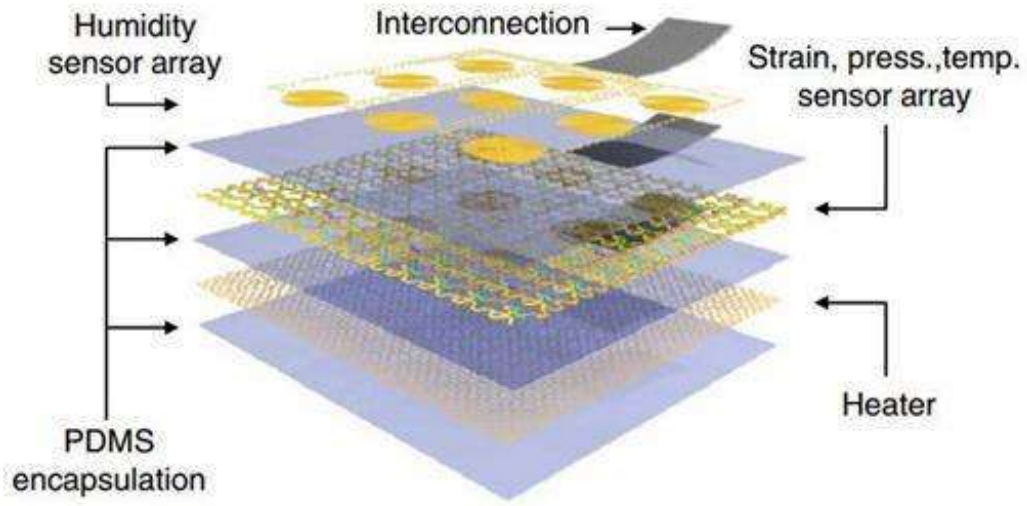
L'un des moyens de se rapprocher et même de dépasser l'homme est de multiplier les capteurs sensoriels. La principale différence entre l'homme et la machine réside dans la portée de ces capteurs. Pour l'homme, la portée est immédiate et ne concerne que ses alentours. Pour les machines, elle peut-être 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.

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 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

[22] 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 !

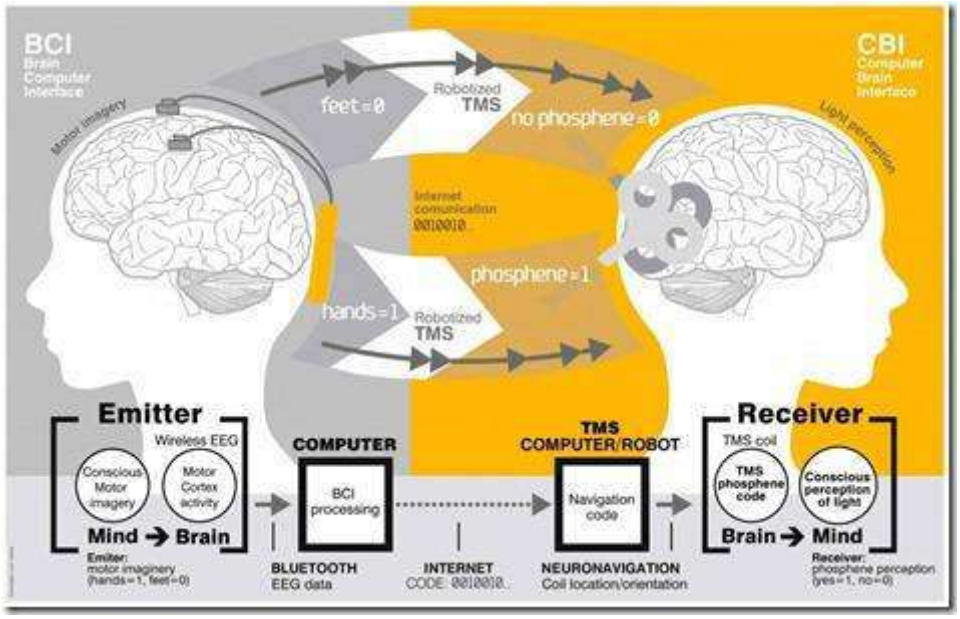
réponse, de vision à 360° et de capacité d'anticipation des mouvements sur la chaussée (piétons, vélos, autres véhicules).



Les capteurs de proximité intégrables à des machines comme les robots progressent même dans leur bio mimétisme. Des prototypes de peau artificielle sensible existent déjà en laboratoire, comme en Corée du Sud (*ci-dessus*, **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.

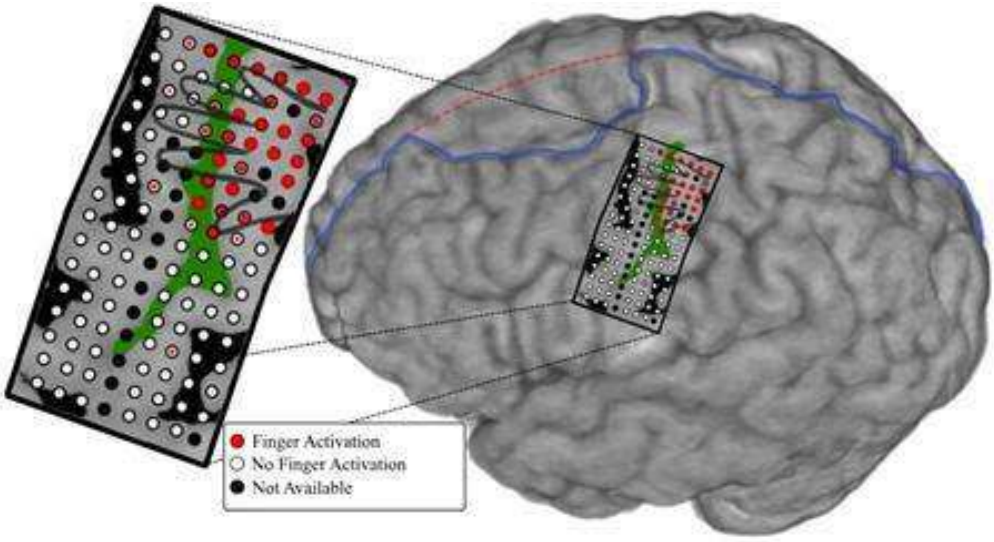
Les capteurs fonctionnent aussi dans l'autre sens : de l'homme vers la machine. Les progrès les plus impressionnants concernent les capteurs cérébraux permettant à l'homme de contrôler des machines, comme pour contrôler un membre artificiel robotisé, une application pouvant restaurer des fonctions mécaniques de personnes handicapées, voire de démultiplier la force de personnes valides, dans les applications militaires ou de BTP. L'homme peut ainsi piloter la machine car la périphérie du cortex cérébral contient les zones où nous commandons nos actions musculaires.

Des **expériences de télépathie** sont également 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.

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 !



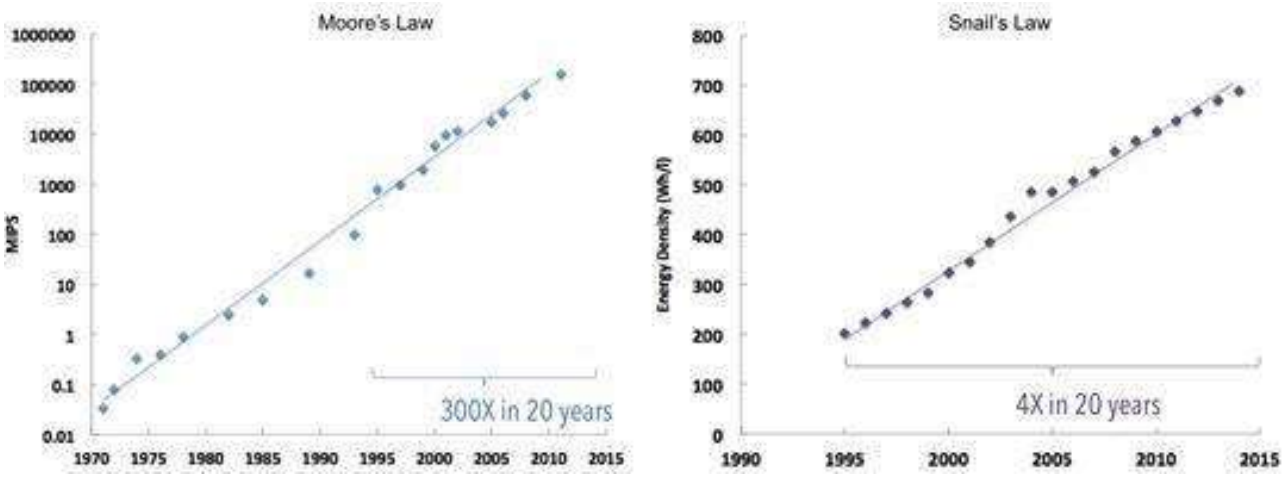
(source de la photo, crédit Guy Hotson)

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 théo-

rie d'atteindre une densité énergétique 20 fois supérieure à celle des batteries actuelles, utilisées dans les véhicules électriques[23].

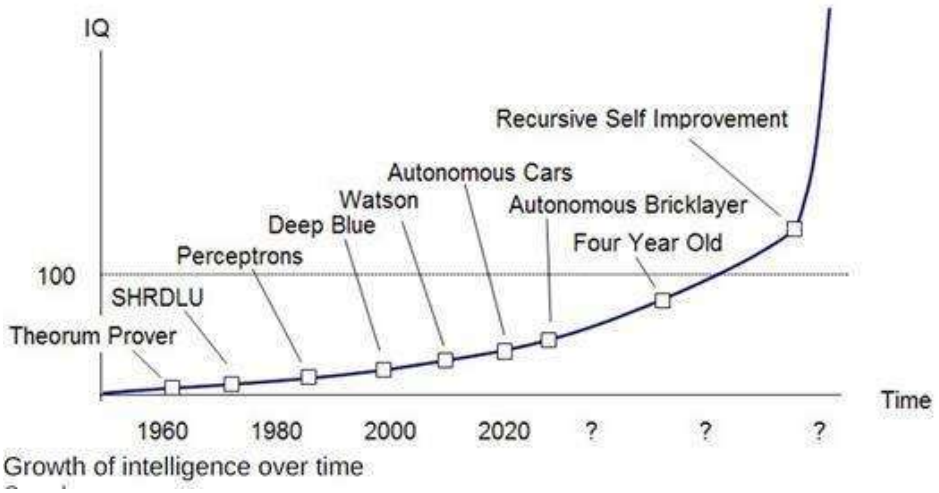
Mais en elles sortent rarement, faute de pouvoir être industrialisées à un coût raisonnable ou de bien fonctionner dans la durée. Parfois, on arrive à une densité énergétique énorme, mais cela ne fonctionne que pour quelques cycles de charge/décharge. Trop injuste !

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.

Sécurité

C'est un sujet évoqué de manière indirecte, au sujet du jour où l'IA dépassera l'intelligence de l'homme et s'auto-multipliera au point de mettre en danger l'espèce humaine. Cela part du principe qu'une intelligence peut se développer à l'infini in-silico. Pourquoi pas, dans certains domaines. Mais c'est faire abstraction d'un point clé : l'intelligence est le fruit, certes, du fonctionnement du cerveau, mais également de l'interaction avec l'environnement et avec les expériences sensorielles.



(schéma tiré de "The artificial intelligence singularity, 2015")

[23] 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.

L'intelligence cumule la capacité à créer des théories expliquant le monde et à des expériences permettant de le vérifier. Parfois, la vérification s'étale sur un demi-siècle à un siècle, comme pour les ondes gravitationnelles ou le 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.

L'IA présente des risques bien plus prosaïques, comme toutes les technologies numériques : dans sa sécurité. Celle 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 ou visant à altérer le comportement de l'IA, 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 machine learning.

Les dangers de l'IA, s'il en existe, sont particulièrement prégnants dans l'interaction entre les machines et le monde extérieur. Un robot n'est pas dangereux s'il tourne en mode virtuel dans une machine. Il peut le devenir s'il tient une arme dans le monde extérieur et qu'il est programmé par des forces maléfiques. Le "kill switch" de l'IA qui permettrait de la déconnecter si elle devenait dangereuse devrait surtout porter sur sa relation avec le monde physique. Les films de science fiction comme Transcendance montrent que rien n'est sûr de ce côté là et que la tendance à tout automatiser peut donner un trop grand contrôle du monde réel aux machines.

L'homme est déjà dépassé par la machine depuis longtemps, d'abord sur la force physique, puis de calcul, puis de mémoire et enfin de traitement. Mais la machine a toujours été pilotée par l'homme. L'IA semble générer des systèmes pérennes dans le temps ad vitam aeternam du fait de processus d'apprentissage qui s'agrègent avec le temps et de la mémoire presque infinie des machines. L'IA serait immortelle. Bon, tant que son stockage ne plante pas ! Un disque dur peut planter à tout bout de champ au bout de cinq ans et un disque SSD actuel ne supporte au mieux que 3000 cycles d'écriture !

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 (cela existe déjà dans le machine learning) mais de surveiller ses évolutions. Il s'agit d'une ONG créée par Elon Musk qui vise à s'assurer que l'IA fasse le bien et pas le mal à l'humanité. Elle est dotée de $1B et doit faire de la recherche. Un peu comme si une organisation était lancée pour rendre le capitalisme responsable[24].

Autre méthode, se rassurer avec **Demystifying Machine Intelligence** de Piero Scaruffi qui 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, l'effet donc de la force brute.

Selon lui, l'homme a toujours cherché une source d'intelligence supérieure, qu'il s'agisse de dieux, de saints ou d'extra-terrestres. La singularité et les fantasmes autour de l'IA seraient 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.

Piero Scaruffi prend aussi la singularité à l'envers en avançant que l'ordinateur pourra dépasser l'homme côté intelligence parce que les technologies rendent l'homme plus bête, 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. 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.

Pour Piero Scaruffi, en tout cas, l'intelligence artificielle est d'ailleurs une mauvaise expression. Il préfère évoquer la notion d'intelligence non humaine. 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 par les chinois qui séquencent 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 !

[24] Cf **OpenAI** dans Wikipedia et **Why you should fear artificial intelligence** paru dans TechCrunch en mars 2016.

Pour **Daniel C. Dennett**, 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.

Et si le plus grand risque était de ne rien faire ? Pour toutes ces technologies et recherches citées dans cet article, est-ce que l'Europe et la France jouent un rôle moteur ? Une bonne part de cette R&D côté hardware est concentrée au CEA. Pour l'industrie, ce n'est pas évident, à part peut-être la R&D en photonique chez Alcatel-Lucent qui même si elle dépend maintenant de Nokia, n'en reste pas moins toujours en France.

Il reste aussi STMicroelectronics qui reste très actif dans les capteurs d'objets connectés. De son côté, la R&D côté logicielle est dense, que ce soit à l'INRIA ou au CNRS.

Reste à savoir quelle "technologie de rupture" sortira de tout cela, et avec une transformation en succès industriel à grande échelle qui passe par de l'investissement, de l'entrepreneuriat et de la prise de risque car de nombreux paris doivent être lancés en parallèle pour n'en réussir que quelques-uns.

Faut-il pour autant lancer une « stratégie industrielle » coordonnée dans l'intelligence artificielle. L'histoire récente a montré que l'Etat devait créer les conditions de l'émergence d'innovation plutôt que chercher à les micro-manager. Donc, il s'agit une fois encore de poursuivre le développement d'une politique qui favoriser l'innovation et les startups. Et aussi une politique qui investit de manière avisée dans les sujets porteurs au niveau de la recherche.

La robotisation en marche des métiers

Prenons maintenant un peu de recul sur la robotisation en marche des métiers liée aux avancées de l'intelligence artificielle vues jusqu'à présent.

Les prévisions sur l'emploi et leurs limites

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.

L'une des conséquences des développements récents de l'intelligence artificielle se matérialise dans ces multiples prédictions d'automatisation et de robotisation de métiers existants. L'économiste John Maynard Keynes s'en faisait déjà l'écho en 1933, avant même que les ordinateurs fassent leur apparition.

Au point que les prévisions vont jusqu'à anticiper la disparition du tiers à **plus de la moitié des emplois salariés** en une vingtaine d'années seulement. En ligne de mire prioritaire : les conducteurs professionnels, remplacés par des véhicules à conduite automatique, les caissières, déjà en cours de remplacement par des automates qui ne font même pas appel à de l'IA, suivis de nombreux métiers de services, notamment dans les professions libérales ainsi que dans la santé.

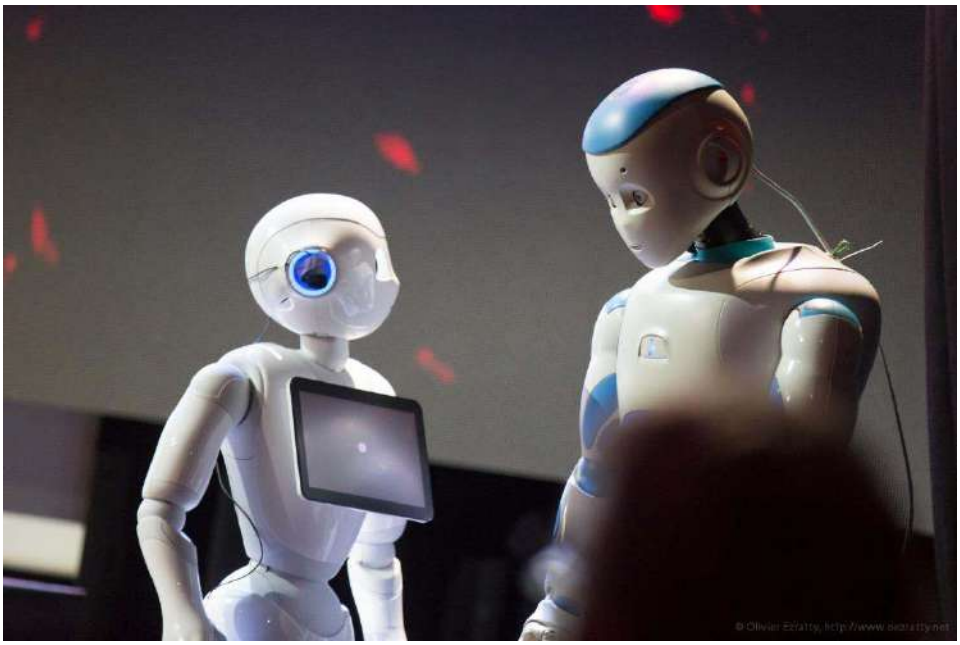
Ces prédictions ne sont pas nouvelles. Certaines d'entre elles datent même des années 1960 ! Au démarrage des précédentes révolutions industrielles, les métiers disparus comme les nouveaux métiers ont rarement été bien anticipés. Pour ce qui est du futur, à vrai dire, on n'en sait pas grand chose. La 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.

Il faut aussi adopter une vue globale de la question. 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 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 des les rendre plus personnalisés.

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**. Des phénomènes de vases communicants peuvent intervenir. Telle disparition entraîne la création d'emploi dans des secteurs connexes voire entièrement différents des métiers disparus.

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.

Ainsi, dans **Les robots veulent déjà nous piquer notre job** d'Emmanuel Ghesquier qui commente une étude d'un certain Moshe Vardi de l'Université Rice du Texas, il est indiqué que "*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 (**photos**) où ils commençaient à être déployés.



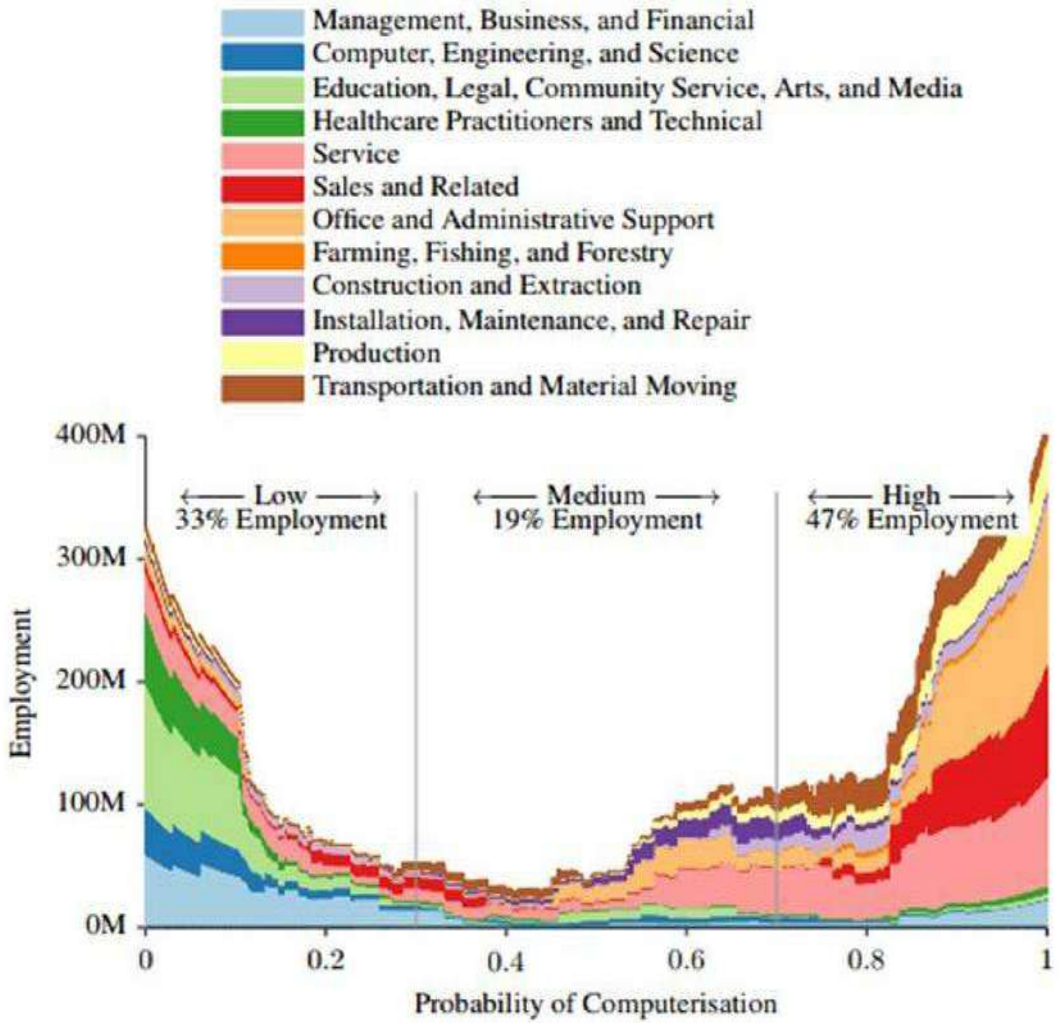
Les robots Pepper et Romeo d'Aldebaran Robotics (groupe Softbank) en apparente discussion, pendant l'événement des 10 ans de Cap Digital à Paris fin mars 2016. En fait, ils ne discutent pas vraiment. Et un shutdown avait bloqué le premier pendant de longues minutes. Work still in progress !

En y regardant de près, l'étude en question est un article publié dans The Conversation, **Are robots taking our jobs**. 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 qui aurait nommé en 2014 un logiciel d'intelligence artificielle à son board dénommé VITAL ! Il sert à identifier les projets les plus prometteurs dans la santé, la spécialité de ce fonds de capital risque. 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. 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 ! Au passage, on ne peut que nommer des personnes physiques dans ces rôles-là, même à Hong Kong [25]!

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. L'innovation a ceci de particulier qu'elle permet aussi la création de métiers nouveaux, qu'elle recompose le paysage économique entier et celui des métiers.



[25] Cf https://en.wikipedia.org/wiki/Deep_Knowledge_Ventures.

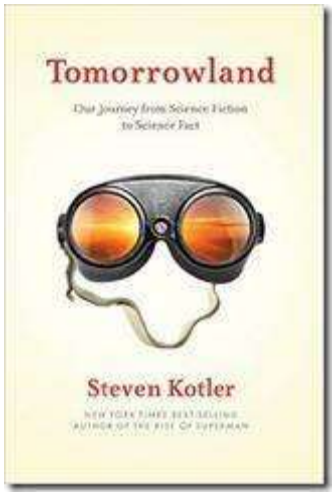
L'une des études les plus sérieuses sur cette question est anglaise : The Future of Employment: How susceptible are jobs to computerisation? Publiée en 2013, elle segmente avec plus de précisions que la moyenne les métiers et leurs risques d'être remplacés par des machines. Le calcul du risque s'appuie 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 (*schéma ci-dessus*).

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 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. Récemment, les agents conversationnels (chatbots) sont revenus au gout du jour, du fait de divers lancements comme chez Facebook.

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**, ce qui ne présage rien de leur capacité à réaliser l'objectif annoncé mais illustre la difficulté à robotiser certains métiers, 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. Nous avons vu dans la partie précédente que la pérennité ou tout du moins la stabilité de la loi de Moore étaient loin d'être évidentes. Les prévisions sont souvent vaguement plausibles mais très vagues côté timing !

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 bien calmée les 5 années suivantes !

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 (sélection directe à l'arrivée) des centres d'appels, les banques en ligne et les assurances. Comme la valeur économique du temps des gens à faible revenu est faible, elle est absorbée en échange de services en théorie plus rapides. C'est un principe également courant dans l'économie collaborative, qu'elle concerne les professionnels (cas de 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.

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 !

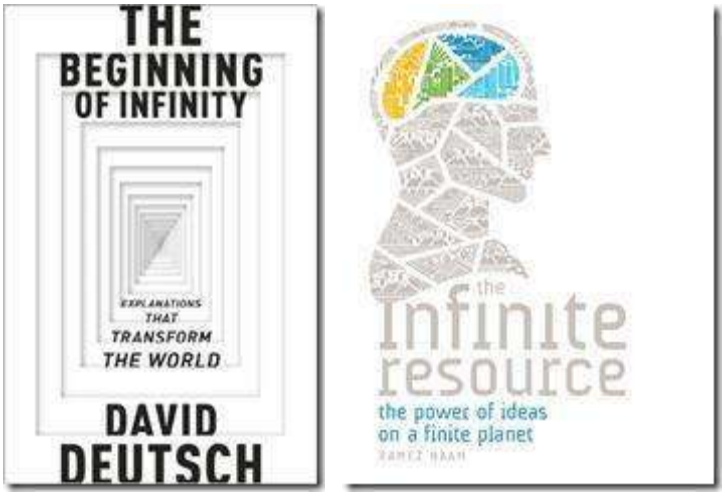
Les ouvrages n'intègrent pas non plus dans leurs scénarios l'épineuse question de la production d'énergie qui est incontournable pour tout développement économique, y compris dans le numérique. L'énergie ? Elle est là, c'est le soleil ! Le seul véritable problème est d'apprendre à la dompter et à la stocker.

Les matières premières ? On en trouvera en fonction des besoins. Ces derniers font évoluer les techniques de recherche et on trouve de tout, y compris dans le domaine des terres rares.

Et la question environnementale qui nous rattrape à grandes enjambées ? Le passage au solaire réglera naturellement une bonne partie du problème sur le long terme !

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 so-

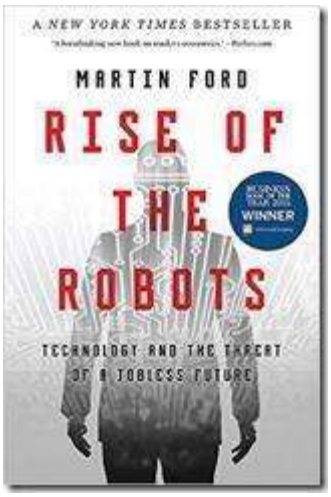
ciétés procrastinatrices et des fainéants : ne nous attaquons pas aux problèmes qui fâchent et attendons que l'IA et la robotique fassent le boulot à notre place ! C'en est presque un éloge du laisser-aller.



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.

Revue de lectures

Aller, c'est parti pour une petite revue éditoriale sur la robotisation des métiers et le futur des emplois !



Rise of the robots and the threat of a jobless future (2016) de Martin Ford est l'ouvrage le mieux documenté de cette série. Il 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 pour l'administration de Lyndon B. Johnson, le successeur de JFK. Ses auteurs s'alarmaient déjà sur risques les risques de l'automatisation, mettant en avant la difficulté de remplacer les emplois supprimés par la modernisation à un rythme suffisamment rapide. Il était très en avance sur son temps, alors que l'informatique n'en était encore qu'à ses balbutiements. Juste avant la sortie du mythique mainframe IBM 360, en 1965, c'est dire !

Aux USA, les 5% des foyers les plus aisées représentaient 27% de la consommation en 1992 et 38% en 2012. Les 80% les moins aisés sont passés de 47% à 39% dans le même temps. Après la crise de 2008, le top 5% avait augmenté ses dépenses de 17% et le reste n'avait fait que rester au niveau de 2008. D'ou l'émergence de business comme Tesla qui cible, pour l'instant, surtout les 5% les plus riches. Les nouvelles entreprises issues du numérique sont automatisées dès le départ et ont moins de salariés. Elles profitent à plein de la productivité issue du numérique. Les exemples un peu éculés et trop généralisant de Whatsapp et Instagram sont mis en avant pour illustrer le point. On nous bassine un peu trop avec les $16B de "valeur" de Whatsapp générés par 55 employés, alors que lorsqu'elle a été acquise par Facebook, cette société n'avait quasiment pas de revenus.

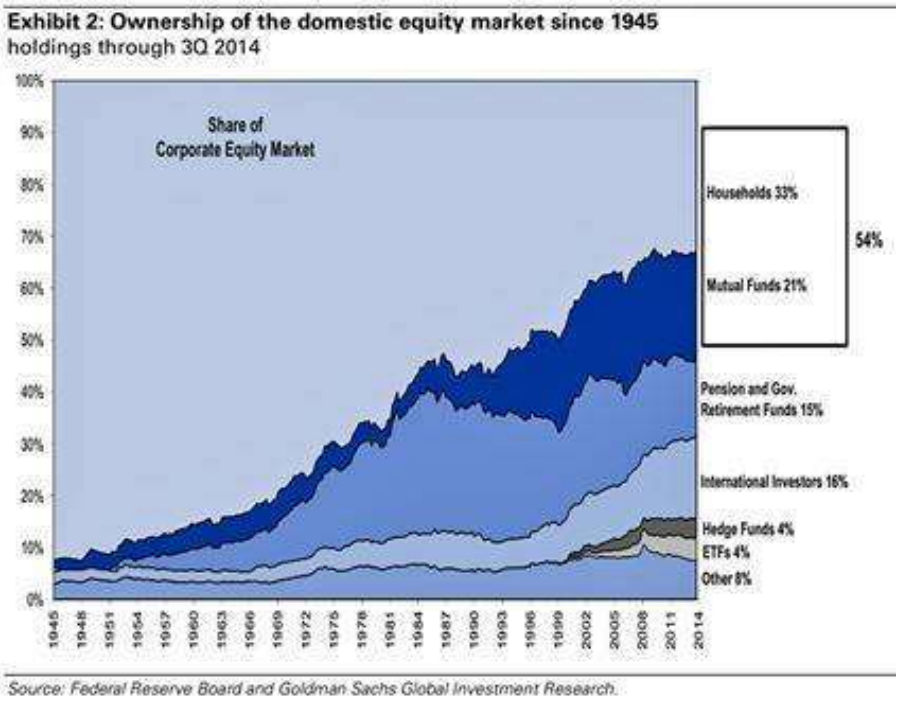
Contrairement à l'après-guerre, les gains de productivité des deux dernières décennies sont allés non pas dans l'augmentation des salaires mais dans la baisse des prix, dans les salaires de métiers techniques qualifiés, et le capital s'orientant vers le financement des nouveaux investissements technologiques. Les technologies sont devenues un facteur d'inégalité au profit des technologues et des détenteurs de capital, tout du moins aux USA. La "finance" réalloue aussi les profits au bénéfice des plus

riches. Plus un pays a un système financier développé, plus grandes seraient les inégalités.

Les profits des grandes entreprises ont augmenté sur 15 ans en proportion du PIB comme indiqué dans le schéma ci-dessus, qui correspond aux données US. Cette réallocation concerne 2,5% du PIB. Je me suis demandé où allaient ces profits (**source** du schéma ci-dessous).

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 !

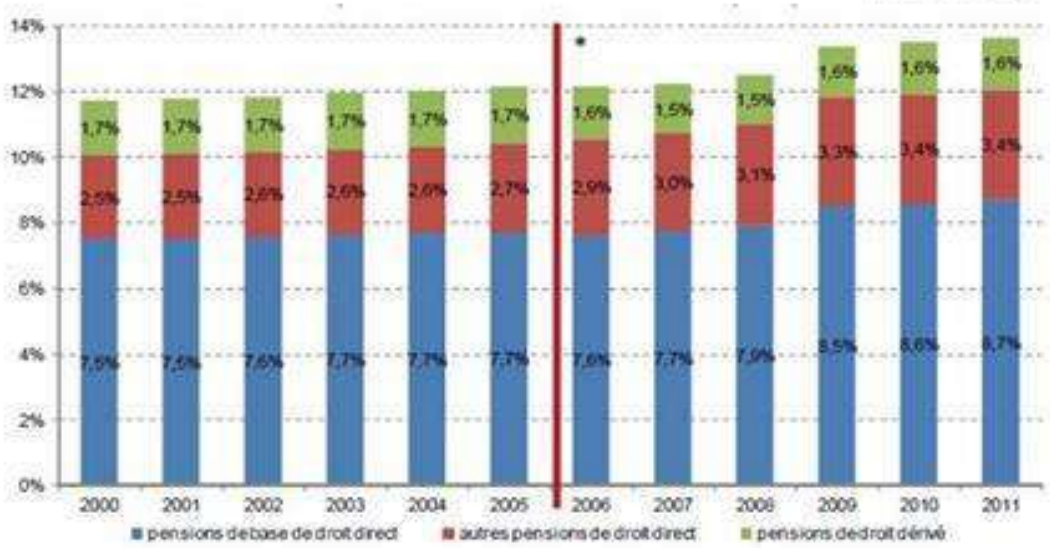
Et si l'explication était donc toute simple : plus la population vieillit, plus les systèmes de retraites par capitalisation ont besoin de financement, donc de profits des grandes entreprises !



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 !

Graphique n° 2 : Evolution de la part des dépenses de retraite dans le PIB

(en % du PIB)



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 type H1B1 pour les cadres qualifiés étrangers imposée par le congrès US créerait une inertie souhaitable pour protéger les emplois non qualifiés. En même temps, elle favorise l'offshore de métiers qualifiés en plus des métiers faiblement qualifiés qui sont déjà externalisés à l'étranger.

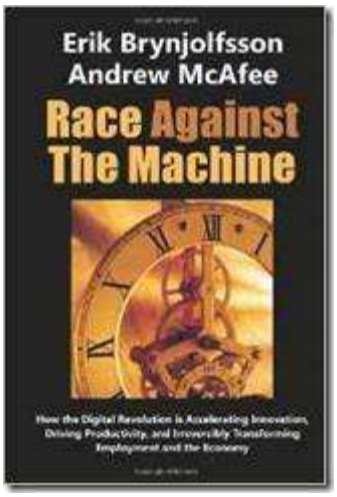
Autre point intéressant, l'auteur fait état des écueils des MOOCs, présentés comme la solution miracle pour l'enseignement. Deux études menées par l'Université de Pennsylvanie en 2013 et qu'il ne faudrait pas forcément généraliser montrent que les résultats d'étudiants ayant suivi des MOOCs étaient moins 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.

Le livre fait aussi état d'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.

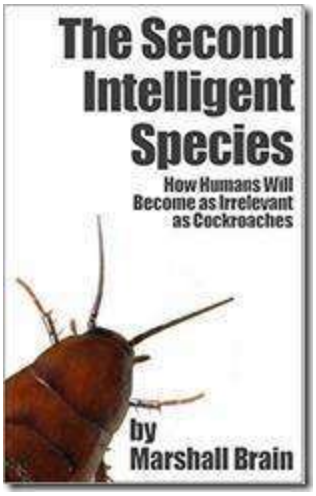
Les deux auteurs, qui sont de la MIT Sloan School of Management, proposent un plan d'action en quatre points qui s'inspire en partie des propositions du **rapport Triple Revolution** de 1964 :

- 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. 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.

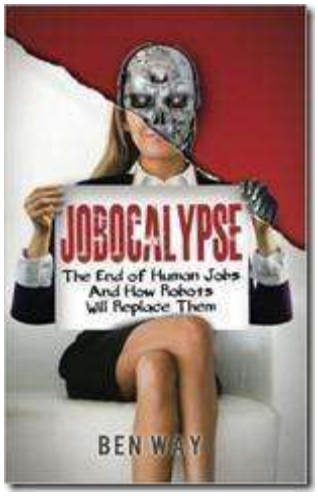
C'en est presque un plan "à la Macron" : favorisons l'entrepreneuriat et tous les problèmes sociétaux se régleront d'eux-mêmes. Un peu trop classique !

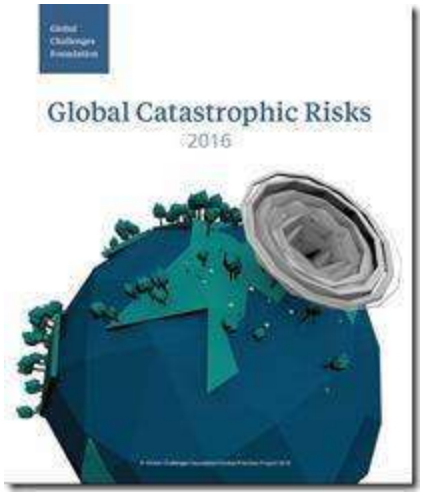


The Second Intelligent Species: How Humans Will Become as Irrelevant as Cockroaches (2015), de Marshall Brain, grossit le trait en annonçant que les scientifiques sont en train de créer une seconde espèce intelligente, les robots et l'IA, qui va nous dépasser et supprimer la majorité des emplois. Les 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.

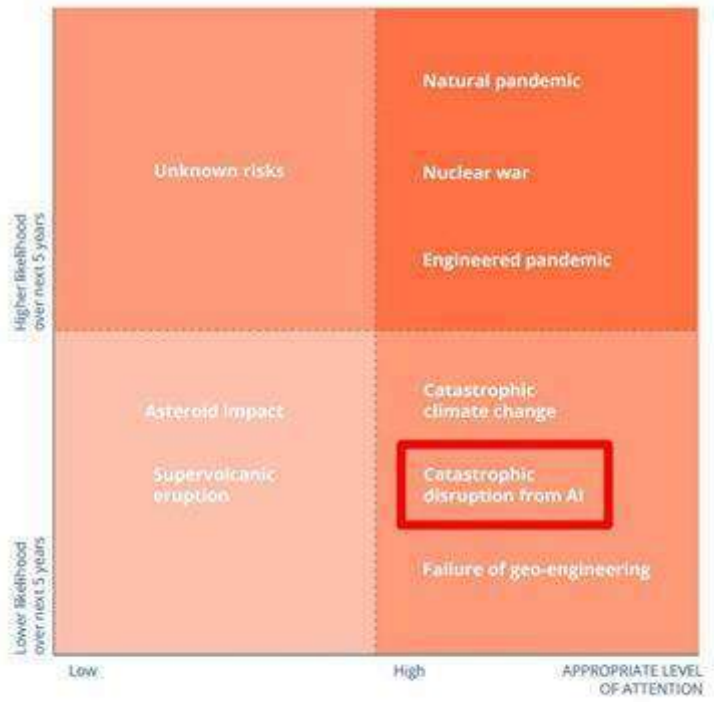
Le reste est de la non-science-fiction, tablant sur une intelligence artificielle qui régulerait les comportements humains néfastes, comme ceux qui affectent l'environnement. Les emplois non qualifiés disparaitraient à la fin des années 2030, ce qui semble un peu rapide au vu de la progression de la 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…



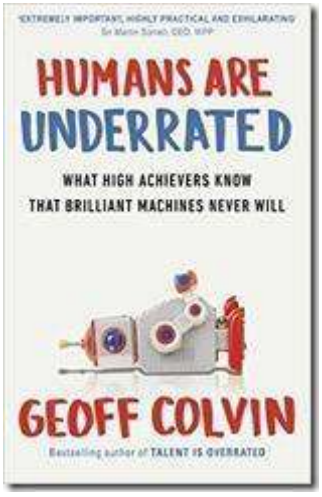


Jobocalypse (2013) de Ben Way, que je n'ai pas lu (désolé…), part du principe que nous sommes *déjà* envahis par les robots et que la disparition d'emplois liée à l'automatisation est une histoire ancienne. Il anticipe que même les métiers les plus qualifiés seront remplacés par des robots car ils s'autoalimenteront. 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 !



Tiré du Global Catastrophic Risks 2016 de la Global Challenges Foundation

Quand au Rapport **Global Catastrophic Risks 2016** de la Global Challenges Foundation, il intègre l'IA dans les risques systémiques que l'humanité et la planète pourraient rencontrer, au même niveau que les conséquences du réchauffement climatique et les pandémies naturelles ou artificielles. Les risques évoqués ne concernent cependant pas les conséquences sur l'emploi mais plutôt la perte de contrôle de l'IA par l'Homme.

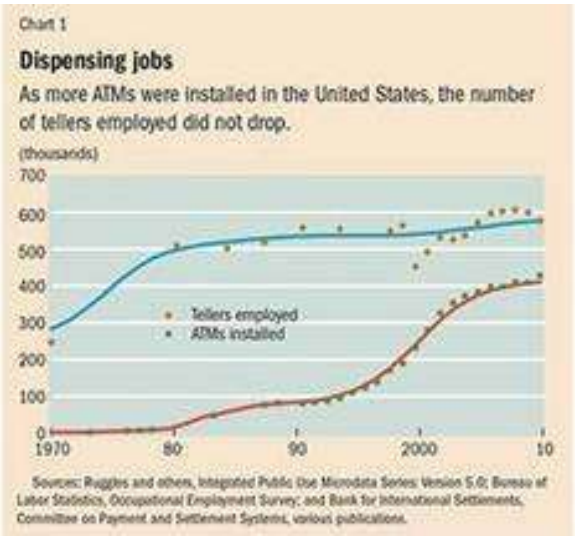


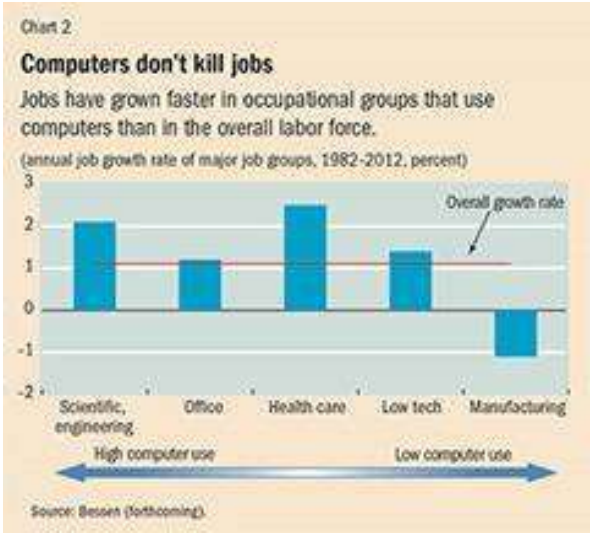
L'étude **AI, Robotics and the future of jobs** du Pew Research Center, parue en 2014, recense de son côté l'avis de divers spécialistes dont certains estiment que la moitié des emplois sont menacés à l'horizon 2025. Ces experts sont très divisés sur la question !

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.

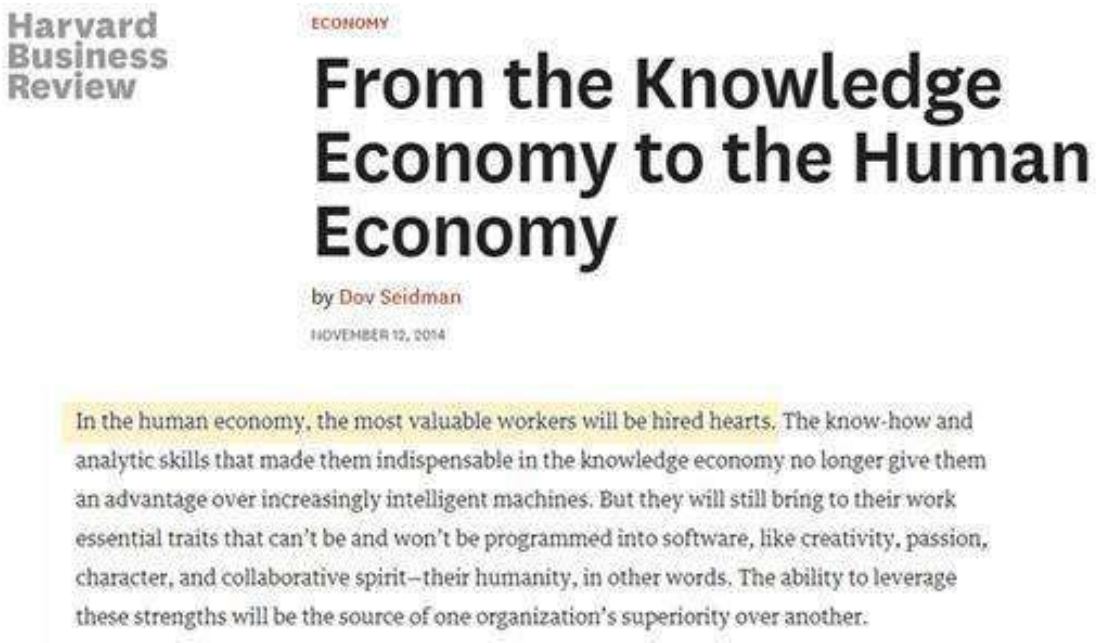
C'est aussi l'avis de Darrell M. West de la Brookings Institution dans **What happens if robots take the jobs** qui prévoit des créations de jobs dans plein de secteurs et des disparitions dans peu de secteurs. On retrouve cette thèse dans **Toil and Technology – Innovative technology is displacing workers to new jobs rather than replacing them entirely** (2015) de James Bessen, 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. Elle était reprise par Dov Seidman dans **Harvard Business Review** en 2014 (*ci-dessous*). C'est une belle conclusion, même si frisant quelque peu l'utopie.


Harvard Business Review

ECONOMY

From the Knowledge Economy to the Human Economy

by Dov Seidman

NOVEMBER 10, 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.

Les incertitudes sur la vitesse de la robotisation

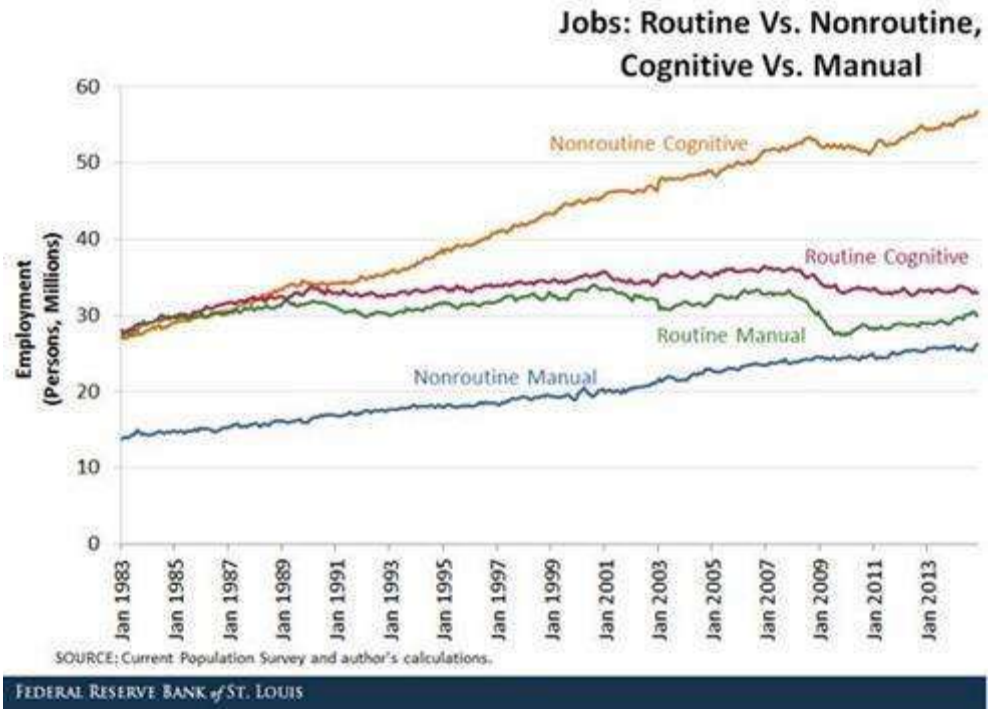
Il y aura bien de l'automatisation et des pertes d'emploi, mais probablement moins que ne l'annoncent les prévisions les plus alarmistes. Les cycles d'innovation sont lents. Les prévisions oublient souvent d'intégrer la dimension économique et sociale.

Pour qu'un métier soit automatisé, même partiellement, il vaut mieux que cela soit intéressant économiquement, et pas seulement pour la performance technologique. Dans certains cas, l'automatisation de métiers pourrait élargir leur marché en les rendant accessibles au plus grand nombre, comme pour les avocats ou de nombreux pans de la santé. Dans d'autres comme pour les traders, l'impact sur l'emploi sera minime et on ne regrettera pas forcément ceux qui occupaient ces métier

L'automatisation à venir est incertaine dans son ampleur, dans sa rapidité, dans les métiers qui seront affectés et dans ses effets induits. Elle produit généralement des transferts de valeur. La France a perdu en un demi-siècle le tiers de ses emplois ! Entre 1946 et la fin du 20eme siècle, l'emploi agricole est passé de 36% à moins de 3% de la population ! Une part de cette transformation a eu lieu pendant les Trente Glorieuses (1945-1975), générant de la croissance dans de nombreux secteurs : l'automobile, le BTP, les biens de consommation courante, le commerce, les loisirs, les médias et la publicité.

Les automatisations à venir ne pourront pas être amorties de la même manière. La "destruction créative" de valeur pourrait avoir du retard à l'allumage. Mais elle libèrera probablement de la créativité dans d'autres domaines, notamment dans les loisirs.

L'automatisation pourrait être plus lente que prévue, notamment dans les métiers les plus difficiles à automatiser. L'ouvrier de chantier n'est pas facile à robotiser de manière générique. Surtout pour les petits chantiers. Mais certains le seront, par exemple via l'impression de béton en 3D qui permet dans certaines conditions de faire l'économie des coffrages.



(**source** du schéma ci-dessus qui montre une évolution à la hausse des métiers cognitifs et non routiniers, donc plus difficiles à automatiser)

Autre point majeur, comment la robotisation interagira avec le réchauffement climatique et ses dégâts ? Est-ce que l'automatisation sera neutre ou positive pour l'environnement ? Sera-t-elle à l'origine de la transformation énergétique nécessaire pour stopper voire inverser le réchauffement climatique ? C'est une question stratégique car le réchauffement climatique pourrait détruire la civilisation actuelle et l'IA comme la singularité pourraient ne rien y faire. L'un des éléments les plus inquiétants est l'impact du réchauffement sur le plancton dans les mers, qui à terme, ne produirait plus une bonne part de l'oxygène nécessaire dans l'atmosphère ! Moins d'oxygène, moins de vie, en tout cas aérobie !

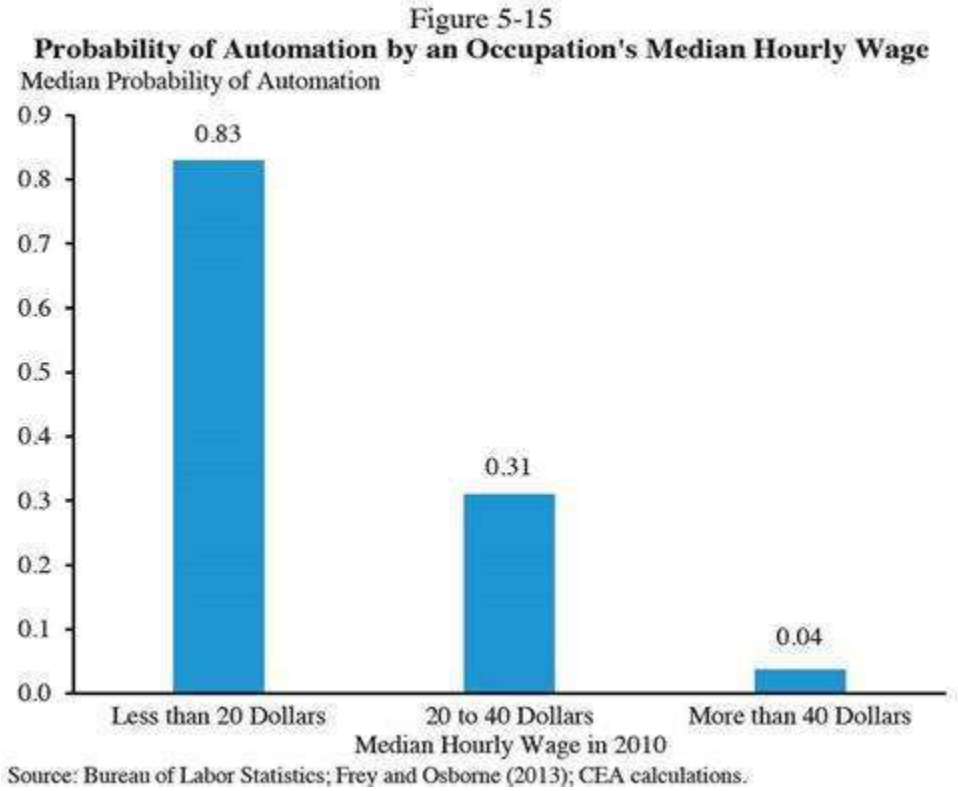
Comment éviter de se faire robotiser

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 l'ubérisation ultime ?

Quelques pistes sont bien connues et déjà cités 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.

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** le montre.



A contrario, il faudra de préférence éviter les métiers répétitifs, routiniers ou à faible degré de créativité et d'initiative et simples d'un point de vue moteur. Ce sont ceux qui présenteraient le plus grand risque d'automatisation.

Le schéma ci-dessus issu du **Rapport Economique du Président US 2016**, déjà cité dans un article précédent 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, 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 dont le contenu émotionnel sera le plus dense, comme expliqué dans **Les 10 compétences clés du monde de demain**.

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.

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.

Et les politiques ?

Un sujet est rarement évoqué dans les ouvrages sur le futur des métiers en relation avec l'IA : quid des organisations politiques et des Etats ? Peuvent-ils 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.

On ne peut qu'aller en conjectures. 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 ou bien du rêve (Obama, Sanders) ou des angoisses (Trump), voir les deux à la fois (H. Clinton). Derrière les deux victoires de Barack Obama se cachent beaucoup de big data et de marketing opérationnel ciblés sur les bonnes audiences.

Que ferait l'IA pour améliorer un tel processus ? Elle collecterait des volumes gigantesques d'informations ouvertes sur ce qui se dit et s'écrit, sur ce que font les électeurs, sur leurs réactions à des discours antérieurs, sur les analyses biométriques (de la captation de pouls avec une montre, des mouvements oculaires avec des capteurs de Tobii, de l'EEG pour la mesure de l'activité cérébrale, etc), sur l'économie ou sur les médias.

Elle les analyserait alors au point de permettre la création de programme politiques appliquant soit la **démagogie ultime** (celle qui fait gagner les élections mais qui est inapplicable ou qui, si appliquée, mène à une catastrophe) soit la **démagogie utile** (celle qui fait à la fois gagner les élections et aller dans un chemin non catastrophique et responsable). Le tout en étant conforme à une idéologie de base d'un parti politique donné, avec son système de valeur (partage, social, économie, croissance,

environnement, fiscalité, justice, école, selon les cas). Voilà de beaux défis d'optimisation sous contraintes !

Political Speech Generation

Valentin Kassarnig
College of Information and Computer Sciences
University of Massachusetts Amherst
vkassarnig@umass.edu

Abstract

In this report we present a system that can generate political speeches for a desired political party. Furthermore, the system allows to specify whether a speech should hold a supportive or opposing opinion. The system relies on a combination of several state-of-the-art NLP methods which are discussed in this report. These include n-grams, Justeson & Katz POS tag filter, recurrent neural networks, and latent Dirichlet allocation. Sequences of words are generated based on probabilities obtained from two underlying models: A language model takes care of the grammatical correctness while a topic model aims for textual consistency. Both models were trained on the Convote dataset which contains transcripts from US congressional floor debates. Furthermore, we present a manual and an automated approach to evaluate the quality of generated speeches. In an experimental evaluation generated speeches have shown very high quality in terms of grammatical correctness and sentence transitions.

Des tentatives de ce genre ont déjà été vaguement lancées. Valentin Kassarnig, chercheur à l'Université Amherst du Massachusetts, a présenté début 2016 un premier **générateur de discours politique** basé sur de l'IA, et qui dépasse les générateurs de pipeau déjà bien connus. Mais cela 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** ! Malheureusement, en politique plus qu'ailleurs, l'adage selon lequel le contraire de l'IA est la bêtise naturelle s'applique parfaitement. Cette dernière est même plutôt efficace électoralement !

Après les élections se pose la question de la gestion. Est-ce que l'IA permettrait de préparer des choix censés mis ensuite dans les mains d'électeurs dans le cadre de démocraties plus participatives ? Est-ce que l'IA permettrait de bâtir des politiques économiques dignes de ce nom ? Est-ce l'IA permet d'intégrer les complexes relations sociales dans la société ? D'anticiper les réactions des citoyens aux nouvelles lois et réglementations, notamment fiscales ? Est-ce qu'elle permettra de gérer les conflits ? Est-ce qu'elle pourrait permettre d'accélérer la justice ? D'éviter les erreurs judiciaires ? De réformer les systèmes de santé au fil de l'eau des progrès technologiques ? Je n'en sais rien. Il n'y a pas beaucoup de chercheurs qui planchent sur ces questions ! **Certains** indiquent toutefois qu'une IA impliquée dans le processus apporterait un peu de rationalité et serait capable de prendre des décisions non basées sur le côté obscur des émotions.

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 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 serait tarder vue l'imagination débridée des scénaristes !

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 !

Epilogue

Nous voici au terme de ce petit voyage dans l'IA. Nous n'avons pas fini d'en entendre parler !

Pour commencer, un grand merci à ceux qui ont commenté les articles à l'origine de cet ebook et m'ont fourni des pistes de réflexion et de lectures. J'avais posé quelques questions clés dans la première partie et les suivants ont permis d'y répondre en grande partie grâce à ces contributions. De nombreuses startups se sont faite connaitre et ont rejoint au fil de l'eau la partie de la série sur les startups françaises.

Les effets de manche sont nombreux derrière les annonces tant dans l'IA que dans la robotique. Une généralisation 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 et le machine learning sont devenus des arguments marketing 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.

L'impression subsiste aussi que les prouesses récentes sont bien plus liées aux progrès dans le matériel que dans les logiciels et les algorithmes. Qui plus est, la puissance brute des machines a tendance à rendre les développeurs moins astucieux dans leur manière d'aborder les problèmes. Mais cette impression vient probablement de la difficulté à bien appréhender la nature même des progrès réalisés dans les algorithmes et procédés techniques de l'IA. Leur vulgarisation est très difficile.

Malgré ces nombreux écueils et la bulle médiatique qui l'accompagne, l'IA fait cependant des progrès réguliers et la vague semble aussi importante que les vagues technologiques précédentes qu'ont été le cloud, le big data ou les objets connectés ou même que les BlockChains, un peu trop présentés en ce moment comme la poudre de perlimpinpin universelle de l'économie. Elle est d'ailleurs reliée à ces vagues car elle en fournit des outils permettant d'en améliorer la performance et la valeur d'usage.

Nous avons aussi vu que la loi de Moore était à la fois un peu enjolivée dans les discours et qu'elle pouvait avoir tendance à ralentir, notamment dans la puissante brute des processeurs. Cela n'empêche pas les chercheurs et entreprises du secteur à redoubler d'efforts pour continuer l'aventure. Ce n'est pas parce que c'est plus difficile qu'avant d'avancer que cela ralentit les innovateurs. Au contraire même ! Les perspectives d'atteindre des "moonshots" les motivent. L'Homme est est insatiable dans sa quête de savoirs et d'innovations.

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 !

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 totale ou partielle de nombreux métiers est déjà en route. 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 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.

Nouveautés à intégrer

Annonce des TPU de Google à Google I/O.

Autres startups qui créent des TPU. Comme celle de Joel Rubino.

https://www.oreilly.com/ideas/the-ai-business-landscape?utm_source=feedburner

http://spectrum.ieee.org/nanoclast/semiconductors/materials/unusual-alloy-brings-magnesiumion-batteries-closer

http://spectrum.ieee.org/nanoclast/semiconductors/materials/bilayer-graphene-could-usher-in-new-tunnel-transistor

https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Toi qui traine souvent dans les coulisses techniques de Roland Garros, je ne sais pas si tu as déjà rencontré Mediawen, cette startup bretonne qui fait de la traduction en temps réel en utilisant, speech to text, machine translation via IBM Watson puis text to speech, en voix de synthèse ou sous titrage. Tu regarderas l'exemple d'ibm/roland garros sur le lien ci-dessous.

Toutes les videos de BIG ont également été sous titrés par Mediawen. Le top management de BPI a été bluffé. Je te montrerai les video (ou tu apparait aussi d'ailleurs).

Je ne sais pas si tu l'as mentionné dans ton rapport sur l'IA, En tout cas on va y investir même s'il y a encore beaucoup d'inconnu. Quel plaisir de savoir que l'on s'embarque vers une telle aventure de la tombée de la barrière des langues; pour eux, la traduction temps réel de très bonne qualité est encore à horizon 5 à 10 ans mais au lieu de rester en laboratoire, ils font du business en disruptant totalement le monde du sous titrage qui est très très artisanal.

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.

Bayésien : technique d'IA s'appuyant sur des modèles probabilistes et statistiques.

Connexionnisme : méthode et techniques de l'IA mettant en œuvre une modélisation à bas niveau à base de réseaux de neurones artificiels.

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.

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.

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.

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.

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, …).

LISP : langage de programmation d'IA utilisé dans les années 80 et 90 et notamment dans la création de systèmes experts.

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.

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.

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).

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.

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 ?

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.

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.

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.

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.

Vie artificielle : simulation de la vie à un niveau d'abstraction arbitraire, via des logiciels.

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 également disponible sous forme d'API en cloud.