La valeur de la donnée

Warith Harchaoui et Laurent Pantanacce

4e Révolution

Juin 2021

English version

Introduction

Chaque révolution industrielle est portée par une force motrice : une matière première, une énergie source, une technologie créatrice qui redéfinit l'économie. Depuis le XIXe siècle, nous pouvons énumérer la vapeur, le charbon, le pétrole, l'électricité, la radio, le transistor, l'informatique et aujourd'hui l'intelligence artificielle (IA). Notre quatrième révolution industrielle est troublante parce que sa denrée est abstraite : la donnée. Pour l'humanité, il s'agit d'un jalon plus marqué qu'un moyen de communication comme la radio ou internet. Au-delà des conséquences industrielles et économiques, il convient d'appréhender l'intelligence artificielle comme un bouleversement comparable à l'agriculture et la sédentarité 10 000 ans avant J.C., l'invention de l'écriture 3 000 ans avant J.C. ou même l'imprimerie au XVe siècle. L'incidence très tangible d'une telle contribution modifie profondément notre rapport au monde. Une machine est aujourd'hui capable de percevoir, traiter et décider sur la base d'informations reçues et l'être humain est alors contraint de remettre en question ses certitudes. Pour les entreprises, presque tous les domaines de notre monde contemporain sont désormais impactés par cette intelligence artificielle. Philosophiquement, elle est l'émergence automatisée de certains aspects du mystère de l'intelligence grâce aux mathématiques appliquées avec des ordinateurs. Concrètement, l'intelligence artificielle est la science qui permet à une machine de prendre des décisions sans que les scénarios aient été exhaustivement explicités par une intervention humaine.

Bien regarder des données, leur accorder la structure qu'elles méritent, les exploiter, les enrichir et les transformer sont les vrais moteurs technologiques de ce qui est bien plus qu'une révolution industrielle contemporaine. À titre de comparaison, il n'y a pas d'autre différence entre le charbon et le diamant que la structure moléculaire : ce n'est « que » l'agencement géométrique des mêmes atomes de carbone répétés partout qui est différent. Pourtant, l'exploitation du charbon a permis un bouleversement d'un retentissement industriel qu'on ne peut raisonnablement pas comparer aux meilleurs amis de Marilyn Monroe même si on peut admirer la grâce de l'artiste, la beauté de la fameuse chanson et l'éclat de ses diamants. Cette même idée de changement de structure,

d'organisation et d'agencement est cruciale pour l'intelligence artificielle et sa matière première que sont les données. Le diamant n'est pas combustible et le charbon est friable alors qu'il ne s'agit que de carbone. Pareillement, les données exigent qu'on les manipule avec soin, une préoccupation et un but en fonction des applications.

Afin d'éviter une exploitation sous-optimale d'une mine de données brutes, le défi continuel consiste à affronter trois questions centrales : Où se trouve la donnée et dans quel état ? Quels sont les outils à disposition pour raffiner la donnée et la rendre utilisable ? Comment se servir économiquement, commercialement et financièrement de cette donnée une fois qu'on la juge bonne à la consommation ?

La donnée dans tous ses états

Le sujet de l'accès aux données n'est pas anodin. L'actualité récente de l'affaire LinkedIn *vs.* HiQ[1] est un bon exemple. En effet, la justice américaine a pris la décision le 9 septembre 2019 (confirmée le 4 juin 2021) de laisser HiQ aspirer les informations rendues publiques par les utilisateurs sur le réseau social professionnel LinkedIn. Le terme consacré en anglais de cette pratique est le *web scraping* et la jurisprudence de cette décision va offrir des nouvelles opportunités « data » pour ce domaine florissant.

D'après le portail Statista[2], la quantité de données informatiques totale dans monde a été multipliée par 20 de 2010 à 2020 pour représenter 47 milliers de milliards de Gigaoctets. On prévoit aussi une nouvelle multiplication par presque 500 entre 2020 et 2035. C'est tellement considérable qu'on peut naïvement se dire qu'il y a sans doute de la valeur dans toutes ces données publiques et gratuites sur internet mais ce n'est malheureusement pas aussi simple. En effet, cette donnée n'est pas toujours utile et vraie mais partons du principe que si. Pour exploiter ces données de la manière la plus pertinente, nous rencontrons quelques obstacles techniques qui réunissent plusieurs métiers pour s'en sortir. D'abord, les données ne sont pas toujours publiquement accessibles et gratuites.

Pour une entreprise, l'effort « Data » est un effort endurant sinon il est vain malheureusement dans ce qu'on observe en pratique. En fait, il s'agit d'une vraie politique volontaire encouragée par la direction sinon quelques travaux seront peut-être faits dans le bon sens mais on retomberait systématiquement sur le problème des données enfermées dans les "silos" informatiques avec des rivalités de politique interne voire de baronnies. Le risque est de payer l'effort de la collecte sans bénéficier des données rendues inaccessibles de fait : enterrées pour des raisons humaines, de sécurité, d'interface informatique et/ou graphique peu conviviale, d'obsolescence, bref de mauvaise gestion.

[1] https://www.reuters.com/technology/us-supreme-court-revives-linkedin-bid-shield-personal-data-2021-06-14
[2] https://tinyurl.com/worldwide-data

Sur internet, les informations peuvent être présentes sous la forme de textes mais aussi d'images, de photographies, de sons et de vidéos. Dans ces cas-là, le recours à l'intelligence artificielle *perceptive* est pertinent. Pour le texte manuscrit ou dactylographié contenu dans les images et les photographies, on utilise l'OCR (*Optical Character Recognition* ou reconnaissance optique de caractères) qui transforme le texte présent sous la forme de pixels en un texte éditable. Pour le son, la transcription audio en texte popularisée pour les dictaphones de médecins retrouve une nouvelle jeunesse avec l'avènement des podcasts (émissions sonores sur le web). Nous avons aussi la reconnaissance d'objets et de personnes dans les photographies avec les progrès spectaculaires du domaine *Computer Vision*. Depuis les années 2010, ces technologies sont devenues vraiment convaincantes pour une utilisation professionnelle, comme un service, une sorte de commodité basique comme l'eau, l'électricité et le gaz à tel point que les expressions *AI is the new electricity* et *AI as a Service* sont devenues des normes. Pour la décennie 2020, on s'attend à d'énormes progrès bien entamés pour la compréhension du texte : émotions, intentions et sens pour ouvrir encore d'autres opportunités.

Extraire l'huile des données

À ce stade, les données sont stockées dans votre système de base de données parce l'effort informatique a été fourni et peut-être aussi l'effort d'IA perceptive. Du *Machine Perception* on peut évoluer vers le *Machine Learning*[3] qui sont évidemment des sciences soeurs sans entrer dans les variations de terminologie liée aux modes changeantes selon les décennies. À présent, nous pouvons faire des étapes exploratoires qu'on appelle les statistiques descriptives qui est un domaine largement sous-estimé alors qu'il est l'un des plus simples et initiaux d'un *processus* IA bien construit. Il s'agit d'interroger les données en calculant des choses simples comme des moyennes, des indicateurs les plus rudimentaires simplement pour avoir les ordres de grandeur des données qu'on manipule. Cette intuition très physicienne[4] de prise en main n'est pas qu'une question de scientifiques, nous nous permettons humblement dans ce document de la recommander aux dirigeants d'entreprises aussi. Il s'agit de s'emparer des enjeux et de se faire une idée au gré de sa propre curiosité intellectuelle.

En intelligence artificielle, sans résumer avec sérieux et en un seul paragraphe un domaine scientifique aussi immense, on peut quand même distinguer :

- l'apprentissage **supervisé** où on a rassemblé des données d'entrée et de sorties désirées (souvent annotées manuellement) pour bâtir un modèle et espérer avoir la bonne sortie sur une entrée que le système n'a jamais vue

[3] https://research.google/research-areas

[4] *The Pleasure of Finding Things Out*, Richard Feynman (publié à titre posthume en 1999)

- l'apprentissage **non-supervisé** où on n'a que les données et on cherche à la décomposer : en dimensions caractéristiques (*dimensionality reduction* pour voir les données en 2D ou 3D), en groupes (*clustering*), en apprenant à les imiter (*génération*) par exemple
- l'apprentissage **par renforcement** où on prend des décisions sans toujours savoir si elles sont immédiatement bonnes. Par exemple, aux échecs, au jeu de Go, en robotique, c'est très utile

Sous l'impulsion d'entreprises comme les GAFAM (Google, Amazon, Facebook, Apple et Microsoft, et Tesla qui rejoint ce groupe en termes de capitalisation boursière) et des fondations privées comme OpenAI, nous accédons à des modèles d'intelligence artificielle disponibles à la demande (moyennant finance) de très bonne qualité parce qu'ils sont appris sur des quantités annotées gigantesques de données. Cela laisse à des entreprises clientes plus petites la possibilité d'offrir des fonctionnalités enrichies par l'IA à leurs clients, comme les *chatbots* (agent conversationnel), la traduction automatique, la reconnaissance de la parole ou d'images, la gestion de la relation client, etc. Concrètement, on peut retenir que les modèles sont prêts à l'emploi à 95 % (c'est juste une façon de parler) et les 5 % restants devant être développés spécifiquement si nécessaire. On peut affirmer que c'est souvent la base d'une bonne première version avant soit de se spécialiser pour un nouveau besoin, soit d'avoir l'ambition consciente de tout refaire chez soi avec les moyens que ça implique. Il nous paraît important de préciser qu'une librairie scientifique toute neuve aux résultats impressionnants n'est pas un produit même quand la tentation de le télécharger gratuitement est grande sans le véritable effort d'adaptation sur les données internes et celles du client sans parler de la mise en production.

Une définition du métier de Data Scientist pourrait être ce fameux *tweet* de Josh Wills (à qui ont doit Spark notamment) en 2012 :

A data scientist is someone
who is better at Statistics than any software engineer
and better at Software Engineering than any statistician

qu'on peut choisir de traduire par :
Un data scientist est quelqu'un
qui est meilleur en mathématiques qu'un informaticien
et meilleur en informatique qu'un mathématicien

Ce métier consiste à pétrir la donnée jusqu'à ce qu'on puisse la cuire dans un format qui a de la valeur. C'est une expérience informatique et mathématique à croiser les données entre elles pour les enrichir. Il s'agit aussi de comprendre que les meilleurs boîtes outils du monde comme scikit-learn[5] ne peuvent rien sur des mauvaises données, mal agencées ou

[5] https://tinyurl.com/scikit-learn-prize

mal vérifiées. Mieux vaut faire simple d'abord parce que tout le monde ne peut se permettre de financer un labo de recherche à la Google en interne (d'ailleurs les chercheurs GAFAM eux-mêmes font d'abord des *baselines* dans leurs papiers avant d'aller plus loin au moins pour comparer les performances). Mettre en production une méthode fiabilisée et contrôlée par des indicateurs de performance (KPI) même rudimentaires pour commencer est une approche saine pour les employés, leur hiérarchie et donc l'entreprise toute entière.

Comme l'explique Ted Benson dans son livre *Automating Paperwork*[6], la raison pour laquelle l'intelligence artificielle du texte (*Natural Language Processing/Understanding*) a un fort impact sur les entreprises de services est que la grande majorité des *processus* peuvent être considérés comme des étapes d'analyse, d'extraction et de conversion de l'information *texte*. Les entrées et les sorties de ces étapes ne sont que des textes (y compris les feuilles de calcul et les bases de données). D'après Gartner (2020), il y a plus d'un milliard de *knowledge workers* depuis 2019, c'est-à-dire des personnes dont le travail reçoit de l'information comme matière première et émet de l'information comme produit. Nous pouvons donc comprendre l'ampleur de cette tendance forte sur les entreprises et nos sociétés. Avec un peu de perspective, on constate que le texte reste encore un grand défi pour les entreprises en tant que support de l'information, probablement parce que le langage est profondément lié à l'intelligence humaine elle-même.

Créer de la valeur avec les données

Maintenant que la *data* est agrégée, torréfiée, enrichie voire synthétisée, il est temps de la valoriser et d'exploiter les filons de valeur qu'elle peut procurer.

Faisons que la valeur réside là où nous ne l'attendons pas !

Les données ne doivent jamais être utilisées à des fins de réassurance. Trop souvent, nous tentons de faire dire aux données ce en quoi nous croyons déjà; sans même tenter de les utiliser pour contrarier nos partis pris, nos idées reçues ou pire encore notre manque de recul en tant que Data Scientist mais aussi en tant que dirigeant d'entreprise. "Laisser les données nous montrer ce que nos esprits ne voient pas ou refusent de voir" est le mot d'ordre, il s'agit de parfois s'empêcher d'imaginer et de s'exercer à confronter notre intuition à la réalité têtue. La pire phrase qui puisse être utilisée dans ce cas là est "d'ailleurs les chiffres ne trompent pas !". Ces quelques mots doivent résonner comme un signal d'alerte; chaque fois qu'ils sont cités directement ou via des synonymes, la méfiance doit naître ! Comme dit le dicton, "Les chiffres sont comme les gens. Si on les torture assez, on peut leur faire dire n'importe quoi"[7]. Les données doivent être là pour changer

[6] https://edwardbenson.com/automating-paperwork

[7] *Nombres en Folie - les divagations du mathématicien fou* (2013) de Didier Hallépée

notre façon de voir les choses plutôt que venir confirmer ce que nous pensons déjà : les données viennent détruire les mythes voire nos barrières mentales !

Il y a quelques années, alors en plein lancement d'un projet d'objet connecté pour gérer les jardins, nous cherchions désespérément à obtenir des statistiques sur les jardins, les piscines, les maisons (principales ou secondaires)… Toutes les études de qualité disponibles sur le marché valaient bien plus que le capital de notre société à peine créée. Contraints à une frugalité de rigueur, il a fallu trouver une manière différente de pouvoir se forger une idée de la taille du marché. C'est alors que m'est venue l'idée de compter les piscines; n'importe quelle vue satellite devient soudainement votre meilleur allié; d'abord un comptage dans le sud de la France puis plus largement. La répartition géographique des piscines est étonnante : quasiment aussi nombreuses au bord de la mer que dans les terres plus reculées; étrangement nombreuses en Allemagne… Tout cela était contre-intuitif. Mais surtout cela forçait à se poser de vraies questions. Le quasi-miracle fut qu'une fois les piscines pointées et comptées, les outils de reverse geocoding permettaient d'obtenir l'adresse des maisons et souvent le numéro de téléphone des occupants : une manne commerciale qu'aucune étude PDF n'aurait su offrir ! Cette anecdote montre à quel point les outils souvent gratuitement disponibles sur internet (Google Maps ici mais en général GitHub ou même des sites interactifs) et qu'en persévérant un chef d'entreprise qui joue à être Data Scientist crée finalement de la valeur. Ne faudrait-il finalement pas révéler le Data Scientist qui réside en chacun de nous ?

L'analyse des données comme un film et non une photographie.

L'expression "it's the journey that matters not the destination"[8] s'applique parfaitement à la donnée. Ce qui crée de la valeur n'est pas un PDF analytique produit dans une finalité unique. La valeur est au contraire créée par le fait d'utiliser des données vivantes via un processus de création/production réutilisable dans le temps afin de suivre des évolutions. Dit autrement, que l'analyse soit faite par un stagiaire ou par un grand cabinet de conseil, elle ne créera jamais autant de valeur que si elle peut être recréée à la volée au regard de nouvelles actualisées. Plus tard, la satisfaction de relancer la même analyse mais sur les données mises à jour est une inspiration créatrice amplifiée par l'expérience de ce qui s'est réellement passé. Dans ce processus d'analyse, un esprit en force de proposition crée donc une valeur irremplaçable par ses allers et retours entre données et réflexion. La pire chose qui pourrait se produire est une excellente analyse statique (éventuellement onéreuse) sur la base de données dont très vite on ne se souviendra ni de la source, ni de la méthodologie qui a servi à la réaliser.

Du partage naît encore plus de valeur.

[8] "c'est le voyage qui importe, pas la destination"

Une fois que la donnée a créé de la valeur pour vous, votre département, votre entreprise, vos clients… posez-vous la question de combien de valeur elle pourrait encore créer ? Souvent, les *Chief Data Officers* trônent sur la donnée comme sur un trésor. Ils ont raison, c'est souvent une partie de l'historique de l'entreprise ! Pourtant la donnée, c'est plutôt un sac de blé. Non utilisée, elle pourrit. Bien utilisée, elle nourrit de nouvelles analyses et crée encore plus de valeur.

Le regretté Hans Rosling[9], Public Health Professor, et surtout immense défenseur de la *data*, ouverte, accessible et partagée martelait dans de nombreuse de ses conférences la chose suivante : "Certains pays ont accepté de rendre publiques leurs données, mais ce dont nous avons réellement besoin c'est, bien sûr, d'une fonction de recherche. Une fonction de recherche qui nous permettrait de copier les données dans un format accessible et de les transmettre au monde entier."

Aujourd'hui, les entreprises devraient avoir la même approche de partage de leur données, elles seraient surprises des trésors forcés de se démasquer derrière les données. Pas seulement au bénéfice des autres mais aussi pour le leur. Partager ses belles données, c'est obtenir que des neurones humains révèlent la valeur qui résidait cachée dans ses propres données. Enfin, célébrons la spécificité de cette denrée *donnée* : si je donne une connaissance à quelqu'un, il s'est enrichi et je ne me suis pas appauvri !

Conclusion

En allant glaner des données pour espérer en extraire de la valeur, des réflexes de bon sens sont de mise : si quelqu'un a collecté ces données, qui est-il ? est-ce que la source est fiable ? est-ce qu'après examen on peut dire qu'un travail de curation a été fait ? Ce sont des vérifications importantes parce qu'elles sont source de biais, ce qui est vénéneux pour des modèles statistiques. De la simple moyenne aux réseaux de neurones de dernière génération en passant par les meilleurs algorithmes des grands spécialistes, la pollution statistique est sournoise. Être au moins conscient que les données ne sont qu'un échantillon statistique aussi grand soit-il peut éviter des erreurs magistrales.

En philosophie des sciences, on doit au statisticien George Box une citation d'un flegme si britannique :

All models are wrong, but some are useful

qui veut dire :

[9] Pour s'ouvrir l'esprit :

- un livre passionnant Hans Rosling *Factfulness : Pourquoi le monde va mieux que vous ne le pensez*, 2019
- toutes les vidéos TED de Hans Rosling : https://tinyurl.com/hrosling-ted

Tous les modèles sont faux, certains sont utiles

Cette idée puissante nous rappelle à quel point le but de la démarche scientifique est crucial. On pose des objectifs si possible tangibles, objectifs et mesurables, sinon l'entreprise perd son énergie créatrice. Contrôler chaque module d'intelligence artificielle est indispensable : en données d'entrée pour vérification et en données de sortie pour mesurer une performance. Nous affirmons que faire autrement revient à se battre contre des moulins à vent de promesses non tenues. Les données ne sont pertinentes et les modèles ne sont utiles que si un but a été défini au préalable. Pour la prédiction d'un phénomène particulier, les données ne seront pas collectées de la même façon. Ignorer malencontreusement les objectifs au nom du téléchargement stakhanoviste d'algorithmes et de données est une source scabreuse de biais et un gouffre d'argent dont personne n'a besoin.

Quand les dangers sont évités, quand les données sont bien ordonnées, leurs utilisations prennent enfin une valeur déterminante pour les entreprises. En domptant sa propre volonté et ses idées reçues, avec la rigueur d'un conservateur de musée, l'esprit aventurier d'un découvreur, le trésor des données s'ouvre ultimement aux entreprises.