Le Big Data, pour faire quoi ? – Agir sans comprendre ? Savoir sans science ?

mardi 3 janvier 2017, par KRIVINE Hubert (Date de rédaction antérieure : 2 janvier 2017).

La presse se fait de plus en plus l'écho du Big Data dans presque tous les domaines de l'activité humaine, depuis les prévisions des élections jusqu'aux ciblages commerciaux par Google en passant par la recherche de nouveaux médicaments, les travaux de linguistique jusqu'aux objets connectés. Pour certains, Big Data évoque Big Brother, tandis que pour d'autres c'est une avancée positive de la connaissance. On trouve là le même dilemme que pour les OGM.

Sommaire

- A-t-on encore besoin de (...)
- Agir sans comprendre ?
- Ce que le Big Data n'est pas
- Les limites du Big Data
- Un futur à découvrir

Précisons d'abord ce qu'on entend par le Big Data. Ce n'est pas seulement le fait de disposer de bases de données gigantesques. Associées au « *deep learning* » [apprentissage profond] ou à l'intelligence artificielle, on assisterait là à la naissance d'une nouvelle de genèse des connaissances.

A-t-on encore besoin de démonstrations ?

Je cite la position extrémiste et hélas populaire d'un Chris Anderson. Le titre de son papier est éloquent (*Science*, 23/06/2008) : *The End of Theory : The Data Deluge Makes the Scientific Method Obsolete*. [Le déluge de données rend la méthode scientifique dépassée], ensuite on lit : *With enough data, the numbers speak for themselves*, [Avec suffisamment de données, les nombres parlent d'eux-mêmes], et plus loin : *Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all*. [La corrélation remplace la causalité, et la science peut même avancer sans modèle cohérent, théorie unifiée voire sans aucune explication mécaniste du tout].

En d'autres termes, correctement analysées, le Big Data devrait abolir la vieille distinction entre corrélation et causalité à laquelle les scientifiques *old fashion* restent attachés. La corrélation relie deux événements qui semblent évoluer systématiquement de pair et quand le premier précède le second, on peut penser qu'il en est à l'origine. La causalité suppose l'un est la raison de l'autre. Les anciens résumaient ainsi cette confusion : *Post hoc, ergo propter hoc* [1]. On connaît la blague qui, partant de la corrélation étroite et bien avérée entre la taille des pieds des élèves et leur niveau en mathématique, en « déduit » que : plus leurs pieds sont grands, meilleures sont leurs connaissances en mathématiques. Corrélation évidente, mais causalité ? [2] Bien entendu, la taille des pieds n'est

pas une cause, elle est simplement un marqueur de l'âge des écoliers !

Ceci dit, plus sérieusement, il est vrai que la mise en évidence d'un très grand nombre de corrélations rendue maintenant possible par le Big Data, même sans lien causal, peut avoir un fort pouvoir prédictif ; quelque fois, il peut même être supérieur à celui de la causalité. Pour poursuivre la blague précédente : donnez-moi la taille des pieds d'un élève et je vous prévoirai en moyenne son niveau en mathématique. Après tout, pour agir, et ne pas se perdre dans de vagues considérations métaphysiques tant prisées par certains intellectuels, n'est-ce pas suffisant ? La réponse est clairement : non !

Agir sans comprendre ?

Voici deux exemples tout aussi caricaturaux, mais tirés de la vraie vie, cette fois-ci.

i) Une étude apparemment savante reproduite sans commentaire par le très sérieux *Quotidien du médecin* du 9 janvier 2015 conclut que « La consommation quotidienne par des collégiens de boissons énergisantes augmenterait de 66 % le risque de développer des symptômes liés à un trouble de déficit de l'attention avec ou sans hyperactivité (TDAH), selon une étude parue aujourd'hui dans « Academic Pediatrics » [3].

Voilà qui n'est pas évident et il aura fallu une grosse base de données pour en arriver là ! Ces données sont incontestables, mais la vérité est que les gamins qui abusent du coca-cola sont statistiquement davantage issus de milieux pauvres. Que la réussite scolaire en dépende n'étonnera alors personne. Donc, pour améliorer la statistique des performances scolaires, rien de tel que de virer les pauvres des établissement [4], mais certainement pas d'interdire les boissons énergisantes. Inversement gageons qu'introduire un distributeur de coca-cola au lycée Henri IV n'en fera pas baisser le niveau ! La température est cause de l'indication du thermomètre et pas l'inverse.

ii) Maintenant avec certains économistes, il y a plus extravagant encore : contrairement aux syndicats qui prônent la réduction du temps de travail contre le chômage, les statistiques prouveraient que c'est dans les pays où les seniors continuent à travailler que le taux de chômage est le plus bas. Je cite :« On pensait alors, bien naïvement, on peut même dire de façon assez paresseuse, qu'en se séparant des seniors, on ferait de la place aux jeunes générations. Alors qu'on le sait désormais, les pays qui ont le taux de chômage le plus faible, sont aussi ceux qui permettent le plus aux seniors de travailler. » (Vincent Giret, du « Monde », France info, 1er juin 2016). Pourquoi ne rit-on pas de Vincent Giret comme on rit des blagues de Coluche qui affirmait que quand on est malade, il ne faut surtout pas aller à l'hôpital parce que la probabilité de mourir dans un lit d'hôpital est bien plus élevée que dans son lit à la maison ! N'est-ce pas un bel exemple de « corrélation remplaçant la causalité » selon Anderson ?

Ce que le Big Data n'est pas

Il faut encore préciser que si la découverte du boson de Higgs ou des ondes gravitationnelles n'a pu avoir lieu qu'en manipulant des millions, voire des milliards de données, ces découvertes sont « classiques » : on savait ce qu'on cherchait. Des milliers de scientifiques et techniciens, à l'aide de dizaines de millions de dollars, etc. ont dû imaginer des dispositifs diaboliquement astucieux pour mettre en évidence des effets extraordinairement faibles, disons imperceptibles, *si on ne les cherchait pas là où la théorie les prévoyait*. C'est, par exemple, pour les ondes gravitationnelles, une variation de longueur de moins d'un milliardième de milliardième de mètre d'un bras

d'interféromètre de 1 km. Aucun Big Data n'aurait pu le trouver. Même chose pour le neutrino prévu en 1930 par Pauli, particule neutre et qui n'interagit pratiquement pas avec la matière et donc très difficile à mettre en évidence (il faudra attendre 1956 !). On voit mal comment le Big Data aussi *big* soit-il, sans guide théorique de recherche aurait permis ces découvertes.

Plus généralement, les découvertes qui ont révolutionné la physique (et peut-être même aussi un peu la philosophie), à savoir la mécanique quantique et la relativité, n'ont rien à voir avec une accumulation forcenée de *data*. Einstein a élaboré la relativité restreinte sur la base des contradiction logiques internes aux équations de Maxwell (celles qui régissent les courants électriques, les champs magnétiques et expliquent les ondes radio) et la relativité générale par les contradictions théoriques apparues au sein de la relativité restreinte. Les données – bien sûr indispensables à la vérification de la théorie – ne viendront qu'ensuite. Ce sont ces théories qui fourniront le point de départ à la fabrication de bien de Big Data et pas l'inverse [5]. En d'autres termes, la sérendipité, c'est-à-dire l'exploitation intelligente d'une découverte fortuite, n'est pas la règle générale des grandes innovations scientifiques.

Les limites du Big Data

Dans l'utilisation de Big Data il y a un principe d'induction, c'est-à-dire de prédiction du futur en fonction de la connaissance - très grande, certes - du passé. On peut donc douter de sa capacité à créer de l'inouï, comme une œuvre scientifique (ou artistique) révolutionnaire. La théorie fait bien mieux : elle a anticipé les ondes électromagnétiques (les ondes radio), le positron (un électron de charge positive), le neutrino, la relativité, les ondes gravitationnelles, le boson de Higgs qui seront découvert ensuite, parce qu'on les cherchait. Contrairement à une idée répandue, les mathématiques sont autre chose qu'une façon de synthétiser un tableau de chiffres [6]. Une équation différentielle, par exemple, est un programme. Et ce programme est si riche que dans bien de cas, on a pas fini de l'exploiter. C'est « la déraisonnable efficacité des mathématiques » de Wigner. Sujet passionnant qui a fait couler beaucoup d'encre et que nous n'aborderons pas.

« Raw data is an oxymoron » [La notion de données brutes est un oxymore à savoir une contradiction dans les termes] a écrit très justement Lisa Gitelman. À part peut-être la distribution en taille ou en poids des conscrits, les données peuvent être socialement biaisées - et pas nécessairement de façon volontaire. Il n'y a qu'à voir les prévisions des élections américaines obtenues à coup de millions de dollars [7]. Et même, sur une plus petite échelle, la prévision des élections françaises. Quand Chris Anderson écrit que « avec suffisamment de données, les nombres parlent d'eux-même », tout est dans le « suffisamment ». En fait, ici, le nombre adéquat de data n'a pu être fourni que par le dépouillement des élections elles-mêmes !

Enfin, la théorie du chaos déterministe (l'effet papillon) nous apprend que même avec une accumulation de données, aussi *big* soit-elle, certains phénomènes resteront imprédictibles au-delà d'un certain horizon. Si un battement d'ailes de papillon à Singapour peut provoquer une tornade en Californie du Nord, comment répertorier les battements d'ailes de tous les papillons, de tous les insectes de tous les oiseaux, etc. ?

Un futur à découvrir

Ces mises en garde partielles (nous n'avons même pas parlé de l'intrusion dans la vie privée) ne doivent pas faire oublier le potentiel de la révolution ouverte par les Big Data. Dans les domaines qui croulent sous le poids de données expérimentales ou observationnelles (comme la biologie, la quête

de nouveaux médicaments, l'astronomie [8], la linguistique, etc.) la recherche portant sur l'exploitation de Big Data est prometteuse. La découverte du Sida dans les années '80 est déjà une première manifestation de son intérêt [9]. Notons que le scandale du Médiator, révélé par la courageuse obstination du docteur Irène Frachon qui a travaillé essentiellement sur des dossiers papier a été le fruit du croisement de corrélations. Disposant de Big Data, elle aurait obtenu beaucoup plus vite ses résultats. Peut-être la démonstration la plus spectaculaire de l'efficacité de Big Data a été la victoire du logiciel AlphaGo, en mars 2016, sur Lee Se-Dol, grand maître sud-coréen du jeu de go. Mais en même temps, peut-on dire qu'on a véritablement « compris » [10] comment la machine a gagné ?

En fait, nous sommes à l'aube de nouveaux développements ; il est donc hasardeux de pronostiquer l'aide de Big Data à la création de théorie, donc de compréhension.

Terminons par quelques remarques sur les effets contradictoires que peut générer l'utilisation de Big Data.

L'aide au diagnostique médical est une question très controversée. Généralement les médecins sont vent debout : même entouré, comme aujourd'hui, de procédés scientifiques de haut niveau, l'établissement du diagnostic relève pour eux davantage de l'art que de la science. Il faut être en condition de voir, d'entendre et de palper le patient. Le traitement à son tour ne dépend pas que de la maladie, mais aussi du malade, etc. Ce à quoi on oppose *i)* que les erreurs médicales faites par des hommes seraient plus fréquentes que celles de la machine (même et peut-être surtout en cas de maladies rares) et *ii)* qu'en cas de pénurie de médecins, mieux vaut des infirmiers rapidement formés aux machines expertes appuyées sur le Big Data que rien du tout. Tous ces arguments se tiennent, mais il ne faudrait pas que le Big Data fassent de nécessité vertu. Déjà aujourd'hui, les médecins passent davantage de temps à lire sur des écrans qu'à ausculter et interroger leurs patients.

Les assureurs tendent à proposer à leurs clients des tarifs personnalisés. Aux clients réputés sans risque (et dont le mode de vie - qu'on doit d'ailleurs être autorisé à suivre - semble raisonnable), des contrats bon marché, aux autres des tarifs plus chers. Les malus-bonus sont un embryon déjà répandu de cette tendance. Assurances comme assurés y gagneraient. Du gagnant-gagnant ? Peut-être au niveau individuel. Mais la logique en est l'abandon du principe de solidarité, voire à terme, l'abandon de l'assurance. Qu'importe à ceux qui, pour leur compagnie, visent seulement le minimum de risques pour le maximum de profits.

Peut-être est-ce hors du sujet, mais l'exploitation de la masse de connaissances (sur le web, mais aussi dans le crâne des citoyens) par Wikipédia est à mon avis un très bon exemple d'une réussite sur le choix de ce qui semble « raisonnable ». Ce système, en perpétuelle évolution, s'autorégule par l'intervention statistique d'une masse d'utilisateurs qui corrigent ses déviances éventuelles, et ce d'autant plus efficacement que le sujet est davantage « cliqué ». À la différence des encyclopédies, ce système fonctionne à peu près sans chef d'orchestre et tout utilisateur peut être rédacteur. Contrairement à ce qu'en attendaient ses détracteurs, cette « ultra démocratie » va finalement générer un corpus de connaissances qui s'adapte très vite aux nouveautés et où le taux d'erreurs est de l'ordre de grandeur de celui des encyclopédies traditionnelles (et nul n'est besoin d'attendre une nouvelle édition pour les voir corrigées). De plus – et peut-être même, surtout - Wikipédia est à la disposition gratuite et permanente de centaines de millions de personnes par le monde qui sont loin d'avoir les moyens de disposer d'une *Encyclopedia Britannica*.

Reste plus généralement une question réelle : ne court-on pas le risque, en éliminant à juste titre le délire, de conforter systématiquement le conformisme ?

Notre conclusion sera banale. Comme toute nouvelle avancée, le Big Data peut être au service du meilleur comme du pire. Tout dépend des finalités de qui l'exploitera. Ce n'est donc pas du Big Data dont il faut avoir peur, mais des sociétés actuelles fondées sur le profit dont il peut décupler le pouvoir. Inversement, on peut rêver qu'ils seront une aide indispensable au fonctionnement démocratique d'une future société libérée de la toute-puissance du marché.

Hubert Krivine

P.-S.

* Article à paraître en février dans la revue Viento Sur.

Notes

[1] « Après cela, donc à cause de cela. »

[2] « Prédire n'est pas expliquer » a écrit le mathématicien René Thom, on pourrait ajouter qu'inversement, comprendre n'est pas nécessairement prédire (voir plus loin « l'effet papillon »).

[3] Référence : May–June, 2015 Volume 15, Issue 3, Pages 297–304. http://www.academicpedsjnl.net/article/S1876-2859%2814%2900425-2/abstract

[4] Remarquons que ça se fait déjà très bien avec la carte scolaire !

[5] Par exemple les mesures de la déviation des rayons lumineux passant près d'un astre massif.

[6] Voir par exemple Bruno Latour dans Science en action, La Découverte 2005, p.574.

[7] Le Monde du 10/11/2016 titrait : « L'élection de Trump et les trois échecs du Big Data électoral », puis écrivait « La campagne de Hillary Clinton, supposée être à la pointe des technologies de ciblage électoral, a échoué à la faire élire. Du côté des médias, les modèles prévisionnistes, eux aussi basés sur le Big Data , n'ont pas su prédire l'issue du scrutin. »

[8] Il y a des centaines de milliards de galaxies composées chacune d'une centaine de milliards d'étoiles !

[9] Le US Centers for Disease Control (CDC) est informé du taux anormalement élevé de maladies rares, la pneumonie Pneumocystis carinii (PPC) et le sarcome de Kaposi chez les jeune hommes homosexuels. On donne d'abord à la maladie le nom de Gay-Related Immune Deficiency (GRID) car elle ne semble toucher que les hommes homosexuels. À la fin de la même année on signale des cas chez les utilisateurs de drogues injectables (voir Wikipedia).

[10] En particulier la machine aura « appris » en jouant des millions de parties contre elle-même.