Enjeux épistémologiques de la science des données

Par Jean-Gabriel GANASCIA
Spécialiste en intelligence artificielle

Après avoir rappelé ce qui fait la singularité des « masses de données », laquelle ne tient pas uniquement à leur volume, mais aussi à leur évolutivité et à leur variabilité, nous montrerons que tant leur accumulation que leur exploitation se sont révélées nécessaires pour les grands acteurs du Web et que cela tient à trois raisons liées à la spécificité des industries du numérique. Nous amorcerons ensuite une réflexion sur la science des données et sur l'opposition entre, d'un côté, ceux qui affirment que désormais les corrélations suffisent et, de l'autre, ceux qui s'en tiennent toujours à l'emploi de modèles et à la fonction épistémologique clef qu'ils occupent dans la démarche scientifique. Nous conclurons sur l'absence actuelle de cadre théorique mathématique de la science des données, tout en évoquant les théories anciennes, celles qui existaient dans les années 1990, et en ouvrant sur des progrès en ce sens.

Données et masses de données

Ancienneté des données

La notion de données n'est pas nouvelle et leur recueil systématique ne l'est pas plus. Comme l'ont déjà rappelé Viktor Mayer-Schonberger et Cukier (2013), dès la plus haute antiquité, au temps des Assyriens, on procédait à des statistiques ; par ailleurs, la Bible évoque le recensement ordonné par le roi David de la population de son royaume en vue de lever l'impôt. Depuis l'avènement des premiers ordinateurs, les données représentées sous forme numérique binaire, comme des suites de 0 et de 1, se manipulent de façon automatique. Cette numérisation ne se limite pas à des données quantifiées ; des textes, des images, des sons, des vidéos, voire des sensations kinesthésiques se traduisent, eux aussi, en flux d'informations et sont traités comme tels. Là encore, rien de nouveau : il y a bientôt quatre-vingts ans, Vannevar Bush, Alan Turing, Norbert Wiener et d'autres encore en eurent l'intuition. Bref, les notions de données, de recueil systématique, de numérisation et de traitement automatique de celles-ci paraissent aujourd'hui bien anciennes.

Innovation et masses de données

L'innovation tient à l'automatisation et à la massification du recueil des données par le truchement de machines. Aujourd'hui, des capteurs robotisés enregistrent et stockent des quantités de données. Des observations de toutes sortes – images, sons, températures – se font sans intervention humaine. À cette automatisation de la saisie s'ajoute la production d'annotations collectives centralisées grâce à l'interconnexion de la population sur le Web. Ce que l'on appelle le *crowdsourcing* – littéralement la « collecte par la foule» –, met à contribution de très nombreuses personnes qui collaborent, consciemment ou inconsciemment, à ce recueil de données. Ainsi, chaque requête faite *via* un moteur de recherche comme Google (et, *a fortiori*, chaque achat) est automatiquement exploitée.

Ordres de grandeur

Pour nous faire une idée du changement quantitatif intervenu, donnons quelques ordres de grandeur en prenant pour référence deux unités bien connues : le livre et la bibliothèque. Si l'on considère, en faisant abstraction des images, qu'un livre compte à peu près un million de caractères typographiques – estimation plutôt généreuse –, et qu'un caractère typographique se code sur un octet, on peut dire qu'il « pèse » un million d'octets d'informations, soit un mégaoctet (1 Mo = 10^6 octets). Le catalogue des livres et imprimés de la Bibliothèque nationale de France comprend à peu près quatorze millions d'ouvrages. En reprenant cette estimation de 10^6 octets pour un livre, le poids de l'ensemble des ouvrages référencés au catalogue de la BNF correspond donc à peu près à 14 x 10^{12} octets, soit quatorze téraoctets (To). Ce chiffre assez conséquent correspondait pour l'homme éclairé du XX^{e} siècle à l'horizon ultime du savoir. Mettons-le en perspective avec les capacités de stockage actuelles : un disque dur externe de vingt téraoctets valant moins de cinq cents euros aujourd'hui, il est donc désormais possible de posséder chez soi l'équivalent numérique de la BNF, pour un coût dérisoire comparé au prix de ce grand bâtiment ! Quant au volume total des données du Web, il a été évalué en 2020 à environ 47 zettaoctets (1 Zo = 10^{21} octets), ce qui fait 47 milliards de téraoctets, c'est-à-dire un peu plus de trois milliards de BNF.

Et ce chiffre s'accroît démesurément. Chaque jour, 500 millions de « gazouillis » sont échangés sur Twitter, ce qui, à raison de 140 caractères par message, engendre environ 25 To de données par an, soit bien

plus que le fonds documentaire de la Bibliothèque nationale de France ! Et encore, nous ne considérons ici que le texte des *tweets*, faisant donc abstraction des images et des sons que l'on y associe souvent. En outre, ceux-ci ne rassemblent qu'une très faible partie de ce que l'on échange sur Internet…

Ces éléments aident à se faire une idée grossière des caractéristiques quantitatives de ce que l'on appelle les *Big Data*. Mais le fait d'emmagasiner de grandes masses d'informations n'est pas leur seule caractéristique. On les définit souvent par la formule des « 3 V » – pour volume, vélocité et variété. Le volume, c'est-à-dire la quantité proprement dite, oscille entre le téraoctet (1 To = 10^{12} octets) et le pétaoctet (1 Po = 10^{15} octets = 1 000 To), à savoir entre un dixième du fonds de la BNF et l'équivalent de cent BNF. La vélocité renvoie au fait que cette masse de données se renouvelle en permanence. Enfin, les données sont variées au sens où elles sont hétérogènes : elles peuvent contenir du texte, des images, des sons, etc. Le texte lui-même peut intégrer différentes langues, divers systèmes d'abréviations, etc.

Particularités de l'économie du numérique

À ce contexte technique de l'automatisation du recueil et à la quantité vertigineuse des données collectées s'ajoutent trois caractéristiques de l'univers numérique qui rendent nécessaire l'accumulation de grandes masses d'informations. Elles tiennent aux spécificités de l'industrie du logiciel, à la vulnérabilité de l'information, sujette aux rumeurs (aux *fake news*) et à l'aspiration à la gratuité qui fut à l'origine de l'essor du Web. Précisons ces trois points pour mieux comprendre la situation actuelle.

Retours d'usage

À la différence des objets techniques traditionnels, comme les voitures, les machines à laver ou les montres, les logiciels apparaissent d'une complexité telle que les ingénieurs qui les conçoivent ne parviennent jamais à les réaliser parfaitement, d'un seul coup, en envisageant toutes les situations dans lesquelles on les emploiera. Pour les aider, sont impliqués les utilisateurs afin qu'ils contribuent à l'amélioration des logiciels en faisant part de leurs impressions et en indiquant les erreurs de fonctionnement qu'ils ont constatées. Ce faisant, les industries du logiciel récupèrent ce que l'on appelle les « retours d'usage », c'est-à-dire tous les dysfonctionnements déplorés par leurs clients, et ce à l'échelle planétaire. Elles amoncellent alors des masses considérables d'informations qu'il leur faut absolument être en mesure de traiter.

Détection de signaux faibles

Alors que les empires industriels d'antan reposaient sur des équipements coûteux – hauts fourneaux, mines, fabriques, usines, etc. –, les industries du numérique recourent essentiellement à la matière grise d'ingénieurs bien formés. L'investissement paraît donc désormais très léger dès lors qu'il repose plus sur des hommes que sur des infrastructures. En contrepoint, de nouveaux acteurs apparaissent rapidement et d'autres disparaissent tout aussi rapidement, car la réputation et la pression sociale jouent un rôle considérable pour assurer la fidélité des utilisateurs : on recourt à tel moteur de recherche et l'on ouvre un compte sur tel réseau social simplement parce que d'autres, en qui nous avons confiance, l'ont fait. Si des informations tendent à discréditer tel ou tel acteur, nous pouvons très rapidement le quitter ayant perdu confiance en lui. De ce fait, toute rumeur qu'on laisse enfler risque potentiellement de déstabiliser les empires les plus puissants. Il est de nombreux exemples de telles déstabilisations, comme celui de la société Dell qui a perdu sa position dominante sur le marché des micro-ordinateurs, parce qu'elle n'a pas su écouter les récriminations des utilisateurs de ses matériels (Jarvis, 2011). En conséquence, il importe aujourd'hui pour un industriel du numérique de récupérer tous les bruits qui le concernent et de les traiter au plus tôt pour mettre en œuvre des stratégies de communication visant à apporter les réponses attendues et à le faire savoir, par exemple en laissant entendre qu'il va changer sa façon de traiter les données personnelles dans le but de prendre soin de ses utilisateurs. Cette auscultation permanente du cyberespace pour parer aux commentaires désobligeants émanant soit de clients sincères, soit d'adversaires résolus, conduit, là encore, à une accumulation considérable de données qu'il faut être en mesure de traiter.

Économie de la gratuité

Enfin, la troisième caractéristique de l'univers du Web vient de la part importante qu'y prend la gratuité. Rappelons qu'avec le Minitel, la France disposait d'une avance technologique dans la mise en place des réseaux de télécommunications. C'est dans ce contexte qu'elle a appris à facturer des services, comme la consultation des banques de données, des horaires SNCF ou des sites de rencontre. L'engouement pour le Web, dans la seconde moitié des années 1990, tint donc non pas à l'absence de services équivalents, puisqu'ils existaient auparavant, du moins en France, mais à leur gratuité : grâce au Web, nous pouvions tout faire, sans rien débourser ! Cependant, pour rentabiliser leurs activités, les acteurs se devaient de trouver des ressources sur un modèle compatible avec la gratuité, ce qui paraissait une gageure. Très tôt, on fit de la publicité. Mais cette publicité tous azimuts restait peu rémunératrice ; elle l'était d'autant moins qu'elle s'adressait à des consommateurs de tous âges et de tous pays, puisque le Web était mondial, et que son accumulation la rendait illisible. Apparut alors le besoin de cibler les annonces. À cette fin, on récupéra toutes les informations utiles pour déterminer automatiquement le profil du consommateur qui sommeille en chacun de nous et, par là même, accroître dans des proportions considérables l'efficacité des messages publicitaires. Là encore, l'accumulation et le traitement de masses de données considérables paraissaient indispensables pour les grands acteurs du Web, et les résultats obtenus à l'aide de techniques d'intelligence artificielle en démontrent aujourd'hui toute l'efficacité.

Science sans causalité, ni modèle

Fin de la théorie ?

Un journaliste très influent, Chris Anderson (2008), proclamait en 2008, dans un article publié par la revue *Wired,* dont il était le directeur de rédaction, « la fin de la théorie » et, par voie de conséquence, l'obsolescence de la méthode scientifique attachée à la preuve et à la compréhension. Son argumentation ne relevait pas vraiment de la démonstration, mais plutôt de l'hyperbole : selon lui, les masses immenses de données collectées permettaient aux grands acteurs du Web, comme Google, de conquérir sans coup férir le monde de la publicité. Face à ce succès sans précédent, d'autres champs de l'activité humaine, en particulier la science, devaient dès lors « tomber » sous la coupe des méthodes de l'intelligence artificielle. De ce fait, on rangera bientôt le raisonnement, la déduction, la logique et la pensée au magasin des oubliettes comme autant de vieilleries inutiles. Il ne faut pas rire de ce constat ! L'article a en effet été cité à maintes reprises comme révélateur d'un tournant épistémologique majeur de la modernité. Et, la preuve étant dépassée, la popularité de cet article atteste de sa plausibilité, et cela seul suffit.

Utilité pratique des corrélations

Les techniques d'apprentissage machine détectent très efficacement des corrélations entre différents paramètres, ce qui rend de grands services dans des secteurs comme la publicité, puisqu'il suffit d'établir des liens entre les comportements des consommateurs et leurs achats pour accroître l'efficacité des annonces publicitaires. De même, ces techniques aident à traduire des textes en constatant que, la plupart du temps, telle locution placée dans tel contexte et se rattachant à telle langue se traduit par telle autre locution dans une autre langue. Elles peuvent aussi inspirer, à titre heuristique, les scientifiques en mettant en évidence telle ou telle hypothèse de travail, par exemple, en suggérant, par la détection de corrélations présentes dans d'immenses quantités d'observations, les effets secondaires de tel médicament ou les facteurs de risques associés à tel ou tel comportement, ou encore en réalisant un diagnostic médical à partir d'une image, comme celui d'un mélanome à partir de photographies de grains de beauté (Esteva *et al*., 2017).

Nécessité de preuves

Cependant, comme le dit Raymond Aron dans sa préface de l'ouvrage de Max Weber intitulé *Le Savant et le Politique*, « la vocation de la science est inconditionnellement la vérité ». Et cette quête de vérité ne saurait se satisfaire de simples corrélations, fussent-elles souvent vérifiées ; les relations causales, les mécanismes explicatifs et, surtout, les preuves demeurent essentiels à la compréhension ; à défaut, nous ne disposerons que de conjectures.

Ainsi, et contrairement à ce qu'affirme Chris Anderson, la notion de preuve n'a jamais été remise en question par le traitement de grandes masses de données. Un exemple en convaincra aisément : il existe une corrélation avérée entre l'application réitérée de crèmes solaires et l'apparition de cancers de la peau. Doit-on pour autant interdire les crèmes solaires ? Sans une expérimentation contrôlée effectuée dans les règles, donc en maîtrisant tous les facteurs, en particulier l'exposition au soleil, une telle conclusion ne pourrait être réfutée, ce qui se révélerait néfaste, car plus personne n'oserait dès lors utiliser des crèmes solaires pour se protéger…

Besoin de théorie

Les faits ne couvrent jamais l'intégralité de l'espace des possibles : le monde brut se donne par l'intermédiaire d'une représentation à travers laquelle on le décrit ; il se donne aussi en fonction de facteurs spécifiques qui restreignent la répartition des observations. À titre d'illustration, si l'on se contentait de collationner des faits bruts, il serait difficile d'imaginer que nous observions autant d'exemples de personnes qui mettent des crèmes solaires sans s'exposer au soleil que de personnes qui en mettent et s'exposent ensuite au soleil… Il existe toujours, qu'on le veuille ou non, que l'on en soit ou non conscient, un biais dans les données. Pour corriger celui-ci, on doit formuler clairement des hypothèses théoriques et rassembler les observations, c'est-à-dire les données, eu égard à ces hypothèses, afin de les valider.

À l'évidence, le traitement de grandes masses de données par les ordinateurs transforme la méthode scientifique. Mais l'affirmation selon laquelle la corrélation supplante la causation n'apparaît pas fondée. Et il en va de même des affirmations selon lesquelles la science avance désormais sans s'appuyer sur des modèles cohérents, sans une quête de théories unifiées et sans recours à des mécanismes explicatifs.

Approches théoriques de la science des données

Alors que, dans le courant des années 1990, l'apprentissage machine reposait sur des théories mathématiques, comme la théorie de l'apprentissage statistique de Vladimir Vapnik (1998) ou l'approche PAC – probably approximately correct – de Leslie Valiant (1984), aujourd'hui l'apprentissage profond (*deep-learning*) ne dispose pas d'autre justification que la constatation empirique d'une remarquable efficacité statistique que jusqu'ici personne n'est parvenu à expliquer clairement en termes mathématiques. À cela s'ajoute l'opacité des conclusions obtenues par les machines entraînées avec de l'apprentissage profond. En effet, il est, la plupart du temps, impossible de comprendre dans chaque situation, ce qui justifie ces conclusions, car celles-ci se présentent comme dérivant de combinaisons pondérées d'un très grand nombre de facteurs. Cette double incapacité à interpréter tant la capacité des machines à apprendre que leurs conclusions fait écho au caractère inexplicable de la science contemporaine qu'évoque Chris Anderson.

Tout cela paraît symptomatique d'une tendance actuelle visant à réduire la vérité à un calcul si complexe que seuls des ordinateurs parviendraient à l'exécuter. En affirmant que nous passons de l'humanisme, c'est-à-dire de la religion de l'humain, au « dataïsme », à savoir la domination de l'homme par des ordinateurs abreuvés de *data*, Noam Yuval Harari (2017), dans son livre *Homo Deus,* révèle un penchant analogue. Dans l'un et l'autre cas, cela revient à se laisser dominer par des machines, en se rangeant systématiquement à leurs conclusions, sans les discuter et, par là même, à renoncer à la raison, à savoir à la capacité de l'homme à appréhender le réel par sa pensée, à renoncer à l'argumentation et au débat face à ces oracles que seraient devenus les ordinateurs programmés à l'aide de techniques d'intelligence artificielle… Or, derrière ces renoncements, plus qu'une conception épistémologique nouvelle, se cache surtout une démission politique face aux pouvoirs des nouveaux acteurs qui maîtrisent ces machines ! Et rien ne dit qu'une théorie mathématique formelle de l'induction, à partir de grandes masses de données, n'adviendra pas dans le futur…

Références bibliographiques

ANDERSON C. (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired*, June 23, https://www.wired.com/2008/06/pb-theory/

ESTEVA A., KUPREL B., NOVOA R. A., KO J., SWETTER S. M., BLAU H. M. & THRUN S. (2017), "Dermatologist-level classification of skin cancer with deep neural networks", *Nature* 542(7639), pp. 115-118.

HARARI Y. N. (2017), *Homo deus : une brève histoire du futur*, traduction de l'anglais de Pierre-Emmanuel Dauzat, Albin Michel.

JARVIS J. (2011), *La méthode Google : que ferait Google à votre place ?*, Pocket.

MAYER-SCHONBERGER V. & CUKIER K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, John Murray, traduction française : *Big Data : la révolution des données est en marche*, Robert Laffon.

VALIANT L. G. (1984), "A Theory of the Learnable", *Communication of the ACM*, vol. 27, n°11, November, pp. 1134-1142.

VAPNIK V. (1998), *Statistical Learning Theory*, New York, Wiley-Interscience.