Paru dans *Le Débat*, n° 207, novembre-décembre 2019, pp. 123-131

Daniel Andler

Nuova scientia ou nouveau style scientifique ?

Vous êtes l'Agence nationale de santé publique. Vous êtes chargée d'évaluer la prévalence de la grippe dans le pays, à l'approche de la prochaine épidémie saisonnière, et d'en estimer le pic. À cet effet, vous mobilisez toute une panoplie de ressources, allant de la biologie fondamentale aux connaissances les plus avancées en virologie, de l'épidémiologie théorique aux études cliniques, démographiques, sociologiques des épidémies de grippe observées et traitées depuis des décennies. Mais vous avez un concurrent : une équipe de Google, sans puiser à aucune de ces ressources, se contente d'inspecter les recherches récentes de termes liés, de près ou de loin, à la grippe. Leur algorithme, baptisé « Google Flu Trends » ou GFT, émet une prédiction qui se révèle exacte, avec deux semaines d'avance sur l'agence nationale. Votre approche fondée sur la théorie est battue par une approche athéorique fondée sur les données pures. Tel fut, exposé dans un article de *Nature* de 2009, le plus spectaculaire succès dont les Big Data – les « données massives » – purent, un bref moment, se prévaloir en matière d'application d'intérêt public.

L'année précédente, Chris Anderson, rédacteur en chef de *Wired,* l'influente revue en ligne, annonçait rien de moins que la fin de la science telle qu'on la connaît, éliminée par les données massives de la même manière que l'épidémiologie de la grippe. Il ne mentionnait pas GFT (mais en connaissait certainement l'existence). Il s'appuyait en revanche, entre autres exemples, sur le logiciel de traduction de Google, « Google Translate », capable, selon lui, de traduire d'une langue quelconque à une autre, sans comprendre ni l'une ni l'autre, par la force des seules statistiques. Ce qu'il ne prévoyait pas, c'est l'effondrement irrémédiable de GFT à partir de 2011 (le logiciel s'est largement trompé à partir de cette année et a été retiré du service peu après). Anderson ne prévoyait pas davantage, à l'inverse, que Google Translate, dont les performances étaient à l'époque en réalité encore assez médiocres, et reposaient encore en partie sur des modèles issus de la linguistique, allait faire, à l'automne 2016, un progrès phénoménal, largement dû à l'abandon de tout recours à des connaissances sur la structure syntaxique des phrases : le nouvel algorithme, GNMT (pour Google Neural Machine Translation), fondé sur l'apprentissage profond *(deep learning)*, s'éloigne encore plus que le précédent de toute compréhension proprement linguistique : il n'exploite plus cette fois, du moins peut-on le croire, que des statistiques raffinées.

Une première leçon à tirer de ces deux exemples est qu'il est imprudent en la matière de prendre pour argent comptant toute proclamation de réussite ou d'échec : il faut y regarder de près. Il faut aussi se garder de concevoir l'avenir à l'image du passé, et croire en particulier que nous avons prise sur les échelles de temps : même lorsque nous parvenons à deviner à peu près où vont les choses, nous sommes bien incapables d'estimer, à un ordre de grandeur près, quand notre prédiction se réalisera : l'échec à dix ans peut n'apparaître cinquante ou cent ans plus tard que comme une péripétie, et un

premier succès être rapidement suivi d'une déconvenue. Il ne s'ensuit pas qu'il soit vain d'évaluer sur la base de ses présupposés les chances d'un programme de recherche ; nous ne sommes pas nécessairement voués à attendre patiemment que les faits nous donnent raison ou tort.

De tout cela Anderson était parfaitement conscient. Comme il l'était de la fragilité de sa thèse générale : il ne l'avançait que pour provoquer le débat, à la manière dont une caricature politique efficace oblige le lecteur à se situer ou à reconsidérer son opinion[1]. Pour caractériser les positions en présence, trois précisions seront utiles. *Primo,* les données massives n'exercent leurs effets que par le canal de différents techniques relevant de l'apprentissage automatique *(machine learning)* : classification, notamment par l'apprentissage « profond » (le *deep learning* qui propulse actuellement l'intelligence artificielle au sommet de l'actualité) ; et *clustering*, c'est-à-dire regroupement d'éléments semblables. Inversement, l'efficacité des algorithmes qui les mobilisent est largement attribuable au fait qu'ils opèrent désormais sur des données massives. Par souci de brièveté, nous parlons seulement de données massives, étant entendu que c'est tout le contexte de leur exploitation qui est impliqué. *Secundo,* ce qui est censé distinguer les données massives des ensembles déjà fort substantiels de données constitués dès le XIX^e^ siècle, et qui n'ont cessé de croître avec le développement de l'informatique, c'est un saut qualitatif dû à la combinaison des « trois V » : leur *v*olume (qui croît exponentiellement et se chiffre aujourd'hui à plusieurs dizaines de zettaoctets, c'est-à-dire de plusieurs dizaines de milliers de milliards de gigaoctets), leur *v*élocité (c'est-à-dire de la rapidité avec laquelle elles sont produites, transmises et traitées), et leur *v*ariété (du fait de la numérisation, toutes sortes de processus engendrent des données très différentes de nature). *Tertio,* les données massives sont mises en œuvre depuis une vingtaine d'années, avec une accélération sensible depuis dix ans, dans la réalisation de tâches pratiques relevant du schéma suivant : une situation particulière d'un certain type étant donnée, en prédire l'évolution, pour décider de l'action la mieux adaptée puis l'exécuter[2] ; pour la commodité, ce domaine de prédilection des données massives sera désigné ici par le sigle non standard PDAS (pour « prédiction / décision / action singulière »).

La position mise en avant par Anderson peut alors être caractérisée succinctement ainsi :

1. Rien ne s'oppose à ce que les méthodes mises en œuvre dans le domaine PDAS ne s'étendent à la science elle-même. Le succès est avéré dans le domaine PDAS, et dans certains domaines scientifiques. Il annonce la fin de la science telle que nous la connaissons.

2. Les données massives sont brutes, c'est-à-dire indépendantes de tout présupposé théorique.

3. Les données massives évitent les difficultés des données statistiques traditionnelles (biais, incertitude…). Elles nous fournissent potentiellement une information complète.

4. En soumettant ces données à une induction probabiliste complexe, indépendante de tout cadre théorique, on peut obtenir toutes les corrélations pertinentes dans un domaine donné, quel qu'il soit.

5. L'ensemble de ces corrélations et d'autres données statistiques constitue le savoir scientifique relatif à ce domaine. Ainsi reconstruite, la science est donc le processus algorithmique par lequel les données massives sont transformées en une connaissance scientifique.

Ces thèses prenaient le contrepied des convictions les plus profondes de la plupart des philosophes des sciences et des scientifiques eux-mêmes. Telle était bien l'intention de l'article, qui, bien qu'il ne se conforme pas aux normes académiques et que son auteur ne soit pas du sérail, a suscité de nombreuses réactions plus ou moins fortement critiques. C'est qu'il ne pouvait être simplement traité comme un fantasme futuriste de plus : de fait, d'une part, il ressuscitait l'un des rêves les plus anciens et les plus persistants de la théorie de la connaissance, d'autre part, il le faisait à la lumière d'une incontestable innovation, dont chacun sent bien qu'elle change les termes du problème. Ce rêve est l'induction radicale, qui permettrait de ne partir que des faits bruts pour ériger, par des moyens purement mathématiques, la meilleure connaissance scientifique possible, sans intervention des spéculations

1. Selon le témoignage fiable de Peter Norvig, cité (à contresens, selon lui) par Anderson : voir « All we want are the facts, ma'am » (consultable en ligne). La tonalité générale du texte le confirme.

2. Ce condamné-*ci* doit-il bénéficier d'une libération conditionnelle ? L'algorithme émet une *prédiction* sur ses chances de récidive, entraînant une *décision* d'accorder, ou non, la libération, et l'*action* correspondante.

produites par l'esprit humain. Se passer de ce dernier, aux capacités limitées et sujet à l'erreur, c'est ce que permettraient les données massives. Au fond, théoriser, c'est deviner, ce qui cesse d'être nécessaire dès lors que l'information est disponible.

Or, non seulement l'objectif d'une science nouvelle, libérée de toute théorie, ne semble pas se rapprocher, mais l'apport des données massives, pour important qu'il soit, est tout autant tributaire de théorie que la science traditionnelle que l'on voudrait congédier. Telle est la leçon la plus générale que l'on peut tirer des principales critiques adressées à la position avancée sur le mode caricatural par Anderson, critiques que nous allons rapidement passer en revue, avant de proposer une interprétation raisonnable de la « disruption » provoquée dans les pratiques scientifiques par les données massives.

Une extrapolation hasardeuse

Commençons par les succès dans le domaine PDAS et le passage à la science. Rappelons d'abord l'impératif de prudence suggéré par notre exemple liminaire. On entend parler plus des succès que des échecs. Les résultats les plus spectaculaires portent sur la reconnaissance d'images, y compris certaines images médicales : les données massives feraient presque aussi bien, voire mieux, que les spécialistes. Même dans ce cas, nous n'avons pas le recul nécessaire : les mêmes algorithmes seront peut-être moins efficaces dans un contexte différent[3]. Dès que l'on s'éloigne de ces cas, il est prématuré de parler de succès autre que commercial : les algorithmes de recommandation gonflent, certes, les commandes d'Amazon et le temps passé sur YouTube, mais les conseils pertinents, dans nos expériences personnelles, sont plutôt rares et ne nous surprennent guère. En matière judiciaire, financière, assurancielle, on ne dispose que de résultats préliminaires. La voiture pleinement autonome ne verra peut-être jamais le jour. Les moteurs de recherche tels Google ou Qwant sont plus impressionnants et d'intérêt plus général, mais ne sont pas autre chose que des machines associatives dont la principale vertu est de dégager les associations pertinentes, en nombre infinitésimal, de la masse gigantesque qu'elles brassent : elles résultent de prouesses en matière de *data science,* elles n'affectent pas la science, du moins directement.

De manière plus générale, quelles raisons avons-nous de penser que les succès, réels ou présumés, en matière de PDAS, s'étendront tôt ou tard à la science ? Les deux entreprises n'ont pas les mêmes fins, du moins en apparence : la science veut d'abord comprendre, quitte à guider l'action dans un deuxième temps ; elle imagine des situations encore irréelles, pour les comprendre puis les réaliser, etc. Anderson reproche à la science classique que les modèles simples qu'elle a produits sont insuffisants[4]. Ce n'est pas une nouvelle, et les modèles complexes qu'on leur substitue progressivement ne doivent rien aux données massives : nous verrons, au contraire, que ces modèles leur opposent une résistance de principe. Le seul exemple concret invoqué par notre auteur, celui du séquençage génomique massif par Craig Venter, s'inscrit dans une tradition bien particulière, celle des essais systématiques, mise en œuvre, par exemple, dans la chimie des colorants allemande au XIX^{e} siècle. Si, en génomique, les données massives – de taille d'ailleurs assez modeste – jouent un rôle crucial, c'est en raison du nombre de combinaisons possibles ; elles n'introduisent aucune nouveauté épistémologique. D'autre part, plus de dix ans après la publication de l'article, on ne perçoit pas de changement dans le paysage scientifique que l'on puisse attribuer aux données massives en tant que telles. Les technologies informatiques et numériques jouent évidemment un rôle considérable et ont profondément modifié le visage de la

3. Une semblable mésaventure est advenue à certains systèmes experts des années 1970.

4. Notons au passage, pour faciliter le lien avec le texte d'Anderson, que les scientifiques parlent aujourd'hui volontiers de « modèles » là où leurs aînés, et les philosophes et historiens des sciences, parlent de « théorie ». Ces deux notions sont pourtant bien distinctes dans leur usage rigoureux, et leurs rapports ont évolué au cours du dernier siècle, mais dans le contexte présent, il faut les considérer comme interchangeables.

science, mais cette évolution est bien antérieure au déluge des *data*[5]. Nous devons chercher ailleurs les raisons d'espérer une science entièrement mue par les données *(data-driven)*, une science émergeant des données, ce qu'on appellera désormais, pour faire bref, une *Google science*.

Le problème des données

Une condition essentielle (c'est la deuxième dans notre liste) pour qu'une telle science soit différente de celle que nous connaissons est que ces données soient brutes, c'est-à-dire ne résultent d'aucune perspective particulière, ne reflètent aucun choix fait par le chercheur, aucun effet des circonstances qui les ont produites. L'idée semble incongrue. Une donnée exprime un fait sous forme d'une proposition, laquelle mobilise certains concepts : le répertoire conceptuel figurant dans une base de données particulière constitue donc un premier choix : certains termes apparaissent, d'autres pas. Un deuxième choix est celui du questionnement auquel les données apportent des réponses : on ne recueille jamais *toutes* les données possibles. Qu'une observation aussi élémentaire semble avoir échappé aux tenants de la *Google science* appelle une explication. Ma conjecture est qu'ils assimilent le monde et les régions du monde qui font l'objet d'une science particulière à des images pixellisées. Chaque région peut faire l'objet d'un relevé de données dans un vocabulaire canonique, comprenant les coordonnées d'un pixel et ses valeurs (0 ou 1, ou une valeur de couleur) : aucun choix, aucune perspective ne sont apparemment intervenus en amont du relevé. On peut se demander s'il est possible de considérer « depuis nulle part » une situation concrète – la pixellisation d'une image n'en est pas un exemple. Mais il suffit de constater qu'en pratique tout ensemble de données mobilisé dans une recherche scientifique répond à un questionnement particulier, formulé dans un vocabulaire restreint.

Ce n'est pas la seule raison pour laquelle les données ne sont jamais aussi brutes que ne l'exigerait une *Google science* idéale. C'est que leur collecte n'est jamais exhaustive ni parfaite. Il est certes théoriquement possible de rassembler *toutes* les données d'une certaine forme, mais en pratique ce n'est jamais le cas. D'une part, des limites sont imposées par le cadre matériel de la collecte, d'autre part, des erreurs sont inévitables : le « nettoyage » des données (c'est le terme de l'art) est une étape incontournable et ne peut se faire que sur la base d'un critère de choix, reflétant une conception préliminaire du domaine étudié. Pour prendre un exemple simple, on éliminera une taille d'humain de dix mètres sur la base de ce que l'on croit savoir de la distribution des tailles dans une population humaine. Un exemple réel est celui du Large Hadron Collider du CERN, qui produit un flux gigantesque de données, dont seule une très faible partie est enregistrée : c'est la théorie qui dicte le choix[6]. Les données nettoyées ne sont pas brutes au sens requis.

Enfin, les données massives résultent de traitements informatiques complexes, qui incorporent des choix théoriques d'un bout à l'autre de la chaîne de production. Elles ne sont pas le simple « reflet » de la réalité : d'autres algorithmes produiraient d'autres données.

Ces difficultés sont connues des statisticiens et d'un bon nombre de scientifiques depuis des lustres. Mais elles ne concerneraient (c'est la troisième clause de notre liste) que les données traditionnelles, non massives. Les données massives y échapperaient justement. Elles n'auraient pas à se soucier des biais d'échantillonnage (car elles seraient exhaustives) ni des erreurs (car la loi des grands nombres les ferait disparaître). Elles renfermeraient, sinon, la connaissance complète du domaine, du moins le minerai dont une telle connaissance pourrait être extraite par un traitement convenable. C'est là une double illusion. La notion d'un catalogue exhaustif des faits relatifs à un domaine donné est, sauf exception, incohérente, du simple fait qu'il existe une infinité de combinaisons possibles des propriétés des éléments, des paires d'éléments, des triplets d'éléments, etc., du domaine. En second lieu, les

5. Voir, par exemple, l'ouvrage de Paul Humphreys, *Extending Ourselves. Computational Science, Empiricism, and Scientific Method* (Oxford UP, 2004, 2e éd. 2007).

6. Exemple emprunté à l'article de Martin Frické, « Big Data and Its Epistemology », *Journal of the Association for Information Science and Technology,* vol. 66, n° 4, 2015.

données massives ne sont pas infinies, et le fait qu'il s'en produit, à chaque seconde, des quantités astronomiques n'entraîne pas que, sur un domaine donné, on dispose d'une quantité virtuellement infinie de données pertinentes. Comme il s'agit généralement de domaines hautement complexes, le nombre de données qui seraient nécessaires pour que l'on soit moralement certain que toutes les configurations possibles sont équitablement représentées dépasse les capacités disponibles dans le monde fini qui est le nôtre.

Le problème de l'induction

Venons-en au nerf de la guerre, l'induction sur la base des données. Les sciences font un usage intensif de statistiques. Ce qui est le plus souvent recherché sont des corrélations entre certaines caractéristiques de membres du domaine considéré – nous sommes désormais quotidiennement inondés d'informations de ce genre : la prévalence de l'asthme est liée à la pollution, etc. Il s'agit pourtant d'opérations délicates, qui conduisent à des erreurs lorsqu'elles sont exécutées sans discernement (ce qui, selon certains auteurs, se produit très souvent[7]). Le problème le plus connu est celui des corrélations accidentelles ou fallacieuses (*spurious* en anglais)[8] : c'est ainsi qu'on a cru montrer une corrélation entre un niveau élevé de la dette relativement au PNB et une croissance faible du PNB – corrélation due à une exclusion contingente d'un groupe de pays[9]. Un problème voisin, moins connu, est celui des comparaisons multiples : deux individus quelconques se ressemblent nécessairement sur certains points ; s'ils ont, par ailleurs, une propriété en commun, comment déterminer, parmi les points de ressemblance, celui qui est responsable de cette propriété ? Il existe différents moyens de surmonter ces difficultés, mais loin qu'elles disparaissent lorsque les données sont massives, elles s'aggravent. En effet, l'un des arguments des défenseurs de la *Google science* est que les données massives nous permettent d'inclure beaucoup de paramètres, en sorte de laisser aux algorithmes le soin de détecter ceux qui sont pertinents. Or, plus il y a de paramètres, plus il y a de comparaisons possibles, et, parmi elles, la proportion des comparaisons sans pertinence augmente. On a montré, de même, que la proportion des corrélations accidentelles augmente également[10]. Ainsi, il vient un moment où trop de données noient les informations susceptibles d'être pertinentes.

Comment les scientifiques compétents parviennent-ils à surmonter les pièges de la statistique et, notamment, à écarter les corrélations accidentelles ? En faisant intervenir leur compréhension du domaine — leur connaissance, si faillible et incomplète qu'elle soit, des mécanismes et des causes à l'œuvre dans le domaine. Or les partisans de la *Google science* prônent par principe une cécité volontaire à l'endroit de ce savoir, qui permettrait, selon eux, d'éviter que les idées du chercheur, les biais de l'esprit humain, les croyances traditionnelles ou encore les conceptions attachées à l'usage des termes ordinaires ne brouillent le message des données – le lecteur aura reconnu là les quatre « idoles » contre lequel l'un des pères de l'inductivisme, Francis Bacon, mettait en garde dans le *Novum Organum* de 1620[11]. Indépendamment de cette préférence de principe pour travailler « les yeux fermés », les données massives nécessitent un partage technique des tâches entre ceux qui produisent les données

7. Voir John Ioannidis, « Why Most Published Research Findings are False », *PLoS Medicine,* vol. 2, n° 8, 2015 ; Donald Berry, « The Difficult and Ubiquitous Problems of Multiplicities », *Pharmaceutical Statistics,* vol. 6, 2007.

8. Il faut distinguer les corrélations accidentelles des corrélations qui, sans refléter une causalité directe, résultent néanmoins de connexions réelles, l'exemple le plus simple étant celui de la cause commune : la baisse du baromètre ne cause pas la tornade, ni l'inverse, mais les deux sont dues à un changement de pression atmosphérique. *« Spurious »* tend malheureusement à désigner indifféremment les deux phénomènes. Les corrélations accidentelles sont une forme d'apophénie, cette tendance à « voir » un *pattern*, une configuration, là où – en un sens – il n'y en a pas : une tête d'auroch dans un rocher, un visage courroucé dans un nuage…

9. L'exemple est donné par Sean Roberts et James Winters (« Linguistic Diversity and Traffic Accidents : Lessons from Statistical Studies of Cultural Traits », *PLoS One,* vol. 8, n° 8, 2013) pour montrer qu'une corrélation accidentelle peut servir d'argument en faveur d'une politique économique néfaste ; la santé fournit maints autres exemples de ce genre.

10. Ce phénomène, assez contre-intuitif, est connu de certains statisticiens depuis quelque temps ; il a été récemment prouvé mathématiquement par Cristian Calude et Giuseppe Longo (*Foundations of Science,* vol. 22, n° 3, 2017).

11. La théorie baconienne de l'induction est cependant plus élaborée que celle de la *Google science*.

– les spécialistes du domaine – et ceux qui les traitent, les *data scientists,* dont les compétences se situent à la frontière de l'informatique et des statistiques. Cette séparation induit sinon une cécité, du moins une restriction sensible des occasions d'éliminer certaines erreurs nées d'un traitement statistique, même irréprochable.

Un autre danger – c'est le dernier que nous mentionnerons – guette la *Google science*. Elle pense détenir sur la science traditionnelle un avantage déterminant s'agissant de l'étude de systèmes complexes. Sans doute est-il nécessaire, lorsqu'on a affaire à un nombre très élevé d'entités et de paramètres potentiellement pertinents, de disposer de beaucoup de données. Le *hic* est que dans beaucoup des systèmes qui nous intéressent, de la physique non linéaire à la finance et à maints phénomènes sociaux, les variables sont fortement corrélées. Il s'ensuit que les lois de probabilité auxquelles elles obéissent ne sont pas gaussiennes (contrairement aux exemples familiers tels que la taille des individus au sein d'une espèce, dont le relevé suit une courbe en cloche). La loi des grands nombres indique qu'en prenant suffisamment de données on réduit l'impact des exceptions, c'est-à-dire l'incertitude. Elle s'applique (presque) toujours, mais la vitesse de convergence, rapide dans le cas gaussien, est beaucoup moindre pour ces systèmes complexes. Il en résulte que même avec beaucoup de données l'incertitude reste forte : si massives qu'elles soient, au regard du problème les données restent rares, et la *Google science* perd l'avantage dont elle se prévalait[12].

Le problème des corrélations

Venons-en, enfin, à la cinquième et dernière thèse : le fruit des données massives est un ensemble de corrélations et autres données statistiques, et c'est en cela que consiste la science – celle de demain, en tout cas. L'objection immédiate est que la science que nous connaissons est infiniment plus qu'une collection de corrélations : elle vise à expliquer les phénomènes et notamment les corrélations. À quoi une réplique possible serait que la notion d'explication a été jugée confuse par certains philosophes, qui ont tenté de montrer qu'elle pouvait être éliminée au profit de la notion de loi, elle-même peut-être réductible à celle de corrélation[13]. Mais cette voie philosophique, d'ailleurs fort étroite, n'est pas celle qu'empruntent les défenseurs de la *Google science* : leur objection est que la science n'est pas une forme immuable, que ses fins ont varié au cours des époques et qu'opposer notre conception actuelle à une vision de ce qu'elle sera demain est incohérent. La *Google science* est présentée comme un nouveau paradigme de la science, en voie de se substituer à la pratique scientifique actuelle[14].

À quoi l'on répondra qu'il est peu probable que la science renonce jamais à deux choses : son désir de compréhension et son pouvoir d'intervention. Comprendre, intervenir font appel en dernière instance aux *mécanismes*. La *Google science* elle-même en dépend : la structure matérielle sur laquelle elle repose, comme tout ce que l'humanité construit, sur notre capacité à mobiliser à bon escient certains mécanismes, qu'il faut avoir au préalable déchiffrés. Or les corrélations, à elles seules, ne nous en livrent pas les clés.

☝

Les critiques qui viennent d'être passées en revue pourraient sembler réduire la *Google science* à une vaste illusion. Elle est pourtant mieux que cela. Voyons pourquoi.

La première chose à souligner est que ces critiques ont la forme générale suivante : une difficulté, un risque d'erreur se présentent qui ne peuvent être résolus par les seules ressources des données

12. Sur ce point, voir l'article détaillé de Sauro Succi et Peter Coveney, « Big Data : The End of the Scientific Method ? », *Philosophical Transactions of the Royal Society A,* vol. 377, 2019.

13. La théorie déductive-nomologique de l'explication a été développée par Carl Hempel. voir, par exemple, Denis Bonnay, « L'explication scientifique », *in* Anouk Barberousse, Denis Bonnay et Mikaël Cozic, *Précis de philosophie des sciences*, Vuibert, 2011.

14. Voir, par exemple, Tony Hey, Stewart Tansley et Kristin Tolle (sous la dir. de), *The Fourth Paradigm : Data-Intensive Scientific Discovery*, Microsoft Research, 2009.

massives (accompagnées, comme toujours, de leur traitement informatico-statistique). Il ne s'ensuit généralement pas que la difficulté soit insurmontable : mais elle nécessite d'autres ressources, qui sont précisément celles de la science telle que nous la connaissons – à la fois des connaissances déjà acquises et des procédés non inductifs, allant de la formulation d'hypothèses à l'introduction de concepts nouveaux, à la conception et l'exécution de nouvelles expériences ou à la construction de modèles mathématiques. Ce ne sont donc pas les données massives que nous sommes invités à rejeter, mais un fantasme auquel elles se prêtent et dont Anderson présente une image aguichante. Elles sont parfois d'un faible secours, il est vrai, mais c'est le cas de toute méthodologie.

À la lumière de ce constat, il semble naturel de ramener les données massives à un outil nouveau qui enrichit la trousse du scientifique, lequel est libre, selon les circonstances, de l'utiliser à la place d'outils plus anciens ou en combinaison avec eux. Mais s'en tenir à cette idée, somme toute assez plate, serait pécher en sens inverse : les données massives sont plus qu'un simple outil. Elles s'inscrivent, en effet, dans une très sensible évolution des pratiques scientifiques. D'une part, le courant inductiviste a repris une telle vigueur que l'on peut parler d'un véritable tournant inductif. D'autre part, l'informatique et plus largement le numérique se sont installés au cœur de la quasi-totalité des disciplines de recherche. Les deux mouvements se renforcent l'un l'autre : le premier donne une légitimité accrue au fruit le plus visible du second, à savoir les données massives, dont inversement la prolifération et la rapidité de transmission et de traitement alimentent concrètement le premier.

Le tournant inductif est également nourri par tout un courant théorique, qui cherche, non sans un certain succès, à ressusciter le programme d'une logique inductive formulé puis développé par Rudolf Carnap à partir des années 1940, et que beaucoup jugeaient définitivement enterré. À ce courant se rattachent les tentatives de formaliser les relations causales, en sorte de permettre dans certains cas d'établir un lien de causalité à partir d'informations de corrélation[15]. Plus largement encore, les efforts pour automatiser le raisonnement scientifique, dans le cadre de l'intelligence artificielle, se poursuivent, et quoique les rapports entre ces différentes tentatives ne soient pas directs, elles se renforcent les unes les autres : au prix d'une simplification, on peut dire que tout succès en matière d'automatisation de la démarche scientifique constitue un suffrage en faveur de l'inductivisme. Il faut reconnaître que les résultats sont controversés et les progrès assez lents ; comme dans d'autres domaines où l'on s'efforce de déployer l'intelligence artificielle, celle-ci trouve sa place aujourd'hui comme adjointe à l'esprit humain, non comme son remplaçant autonome. Il n'empêche que l'inductivisme et les données massives, dont l'intelligence artificielle ne constitue qu'un allié secondaire, ont le vent en poupe.

Ce n'est pas qu'une affaire de mode. Les données massives introduisent un troisième terme entre une science mue par les données *(data-driven)* et une science mue par la théorie *(theory-driven*[16]*)*. Elles font advenir une science *informée* par la théorie, dans laquelle l'activité exploratoire se déploie librement à l'intérieur d'un périmètre défini par une perspective théorique générale : à l'échelle locale, elle laisse jouer, sur un mode quasi perceptif, la recherche statistique de *patterns* susceptibles d'être significatifs et dont l'ensemble constitue une sorte de proto-théorie. Ainsi, pour qui s'intéresse à un mouvement social tel que les Gilets jaunes, les sciences politiques, la sociologie, la démographie, l'histoire fournissent un cadre général fixant en particulier le répertoire conceptuel de l'enquête scientifique. Les données massives, à l'intérieur de ce cadre, pourraient mettre au jour des corrélations restées jusque-là invisibles[17]. Ce genre de résultat ne constitue qu'une étape : il reste à s'assurer de la valeur de ces corrélations, puis à les intégrer dans un schéma explicatif. Mais c'est une étape qui n'existait pas jusqu'à présent, consistant en une heuristique d'une grande puissance.

---

15. L'auteur clé est ici Judea Pearl : voir son *Causality. Models, Reasoning, and Inference*, Cambridge UP, 2e éd. (2000) et *The Book of Why. The New Science of Cause and Effect*, Penguin Books, 2018.

16. On dit aussi *« hypothesis-driven »*, ce qui renvoie en français à la notion de méthode hypothético-déductive.

17. Il s'agit d'un exemple d'école. À ma connaissance, rien de tel n'a été entrepris, et l'on peut douter qu'existe aujourd'hui la méthodologie et les moyens nécessaires pour ce type de questions. Rob Kitchin (« Big Data, New Epistemologies and Paradigm Shifts », *Big Data & Society*, 2014) conjecture que les données massives ne joueront jamais dans les sciences sociales qu'un rôle limité.

Il semble bien, en tout cas, qu'un nouveau secteur se soit adjoint à ce qui compte comme de la bonne science, comme le dit avec finesse l'une des meilleures spécialistes de la question, Sabina Leonelli[18]. Ce n'est ni une science nouvelle ni un simple outil. Il faut peut-être y voir quelque chose comme un nouveau style scientifique[19], comprenant un mode de raisonnement, des objets spécifiques et un critère de vérité ou d'acceptabilité propre. Il n'échappe pas pour autant au critère général de rationalité scientifique : sa place reste conditionnée par sa capacité à s'adjoindre aux autres styles admis aujourd'hui pour contribuer à l'entreprise scientifique. Sous nos yeux se forge peut-être une nouvelle alliance.

Daniel Andler.

18. Sabina Leonelli, « Making Sense of Data-Driven Research in the Biological and Biomedical Sciences », *Studies in the History and the Philosophy of the Biological and Biomedical Sciences : Part C,* vol. 43, n° 1, 2012.

19. Cette notion a été longuement explorée par Ian Hacking (voir « Language, Truth and Reason », *in* M. Hollis & S. Lukes, *Rationality and Relativism* [1982], et l'ouvrage *Scientific Reason* [2009]). Elle n'est invoquée ici qu'à titre de suggestion. Une question intéressante serait de savoir dans quelle mesure les données massives remplissent les conditions proposées par Hacking.