Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ?

Fidelia Ibekwe-Sanjuan

Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ?

Fidelia Ibekwe-SanJuan
ELICO EA 4147 – Université Lyon 3

Les dernières « avancées » en matière des Technologies de l'information et de la communication (TIC) ont accéléré la virtualisation de nombreux secteurs d'activité. Le Big Data, le Cloud computing, l'Open Data et le web participatif entraînent des bouleversements importants en science et en société. Un des effets qui suscite de l'inquiétude est le recours croissant aux algorithmes de traitement des données massives (Big data) comme mode de pilotage des affaires. Le Big data a normalisé et entériné un certains nombre de paradoxes qui méritent que l'on s'y arrête pour en expliciter toutes les implications non seulement pour la science mais aussi pour la société.

1. L'avènement du 4ème paradigme scientifique : vers le Big e-Science ?

La société est confrontée depuis quelques années au phénomène du Big Data, à savoir la disponibilité des données tellement massives qu'il faut des algorithmes et des machines puissantes, souvent distribuées dans le *cloud*, pour les traiter. Attal Butte (2014) estime à quatre zettaoctets le volume de données produit chaque année (1 zetta-octet = 10^{21}). Elles ont besoin d'être analysées pour assurer le progrès de la science. Or, l'analyse et l'exploration de ces grandes masses de données dépassent très clairement les capacités humaines, d'où le recours aux algorithmes informatiques, seuls capables de les explorer.

Jim Gray (2009) estime que ces grandes masses de données représentent le 4ème paradigme de la science. La science aurait d'abord connu des siècles d'empirisme qui constitue son 1er paradigme. Ensuite, elle a connu des siècles de théories visant à modéliser et à généraliser les observations empiriques. Cela a débouché sur les principales lois de la physique (lois de Kepler, de Newton, équations de Maxwell) ce qui constitue le 2ème paradigme. Les problèmes modélisés par ces lois étant devenus trop complexes pour être observés directement par l'homme, il a fallu recourir aux machines capables de faire des simulations, constituant ainsi le 3ème paradigme qui a conduit la science jusqu'à la deuxième moitié du dernier millénium. Maintenant, ces simulations et les appareils de capture génèrent des données tellement massives que les scientifiques ne peuvent plus les analyser directement, à l'oeil nu. Ces données massives constituent donc le 4ème paradigme scientifique selon Gray (2009). Le bouleversement que le phénomène de données massives devrait occasionner dans la science est diversement appelé « *data-intensive science* », « *e-science », « cyberscholarship »* ou encore *« data-driven scholarship*» (Atkins *et al.* 2003, Hey 2006). Certaines disciplines connaissent déjà des projets de e-Science. En physique, le plus connu est le projet d'accélérateur de particules, le 'Large Hadron Collider[1]' (LHC) porté par un consortium mondial de plus de 100 laboratoires et de plus de 10 000 scientifiques. La découverte récente en 2013 de la particule boson de Higgs (*Higgs boson*) est un des résultats les plus importants du projet LHC. Les expérimentations du LHC délivrent des données 40 millions de fois par seconde, soit 150 millions de pétaoctets de données par an ou 500 exaoctets par jour.

[1] http://home.web.cern.ch/topics/large-hadron-collider

Le projet SDSS (Sloan Digital Sky Survey[2]) est l'un des plus ambitieux projet en astronomie. Il a pour objectif de scruter un quart du ciel afin d'identifier puis de répertorier des objets célestes. Depuis 2000, le télescope SDSS a collecté plus de 140 téraoctets de données faites d'images 3D. Son successeur, le 'Large Synoptic Survey Telescope' qui le remplacera en 2016 est censé amasser ce même volume de données tous les cinq jours ! Plus de 230 millions d'objets célestes ont été détectés par SDSS. Ces objets doivent être identifiés et classés. Or, dépassés par la dimension de la tâche, les scientifiques en astronomie ont mis en place une expérience de science participative ou de *crowdsourcing* via le projet Galaxy Zoo[3] lequel sollicite le public-amateur pour aider les scientifiques à accomplir cette immense tâche de classification. En sciences humaines et sociales (SHS), des initiatives de e-Science, plus connues sous le nom de '*Digital Humanities'* ou *'Humanités Numériques'* se multiplient. L'objectif est également de sensibiliser les chercheurs en SHS de l'utilité de recourir aux méthodes d'exploration de grandes masses de données pour sonder les collections de données disponibles dans ses disciplines. Le défi international '*Digging into Data[4]'* symbolise les efforts des agences mondiales de financement de la recherche pour arrimer les pratiques de recherche en SHS à celles de e-Science déjà en vigueur dans les disciplines des sciences. Cette communication revient sur les implications et enjeux du pilotage de la science et de la société par les algorithmes du Big Data. Nous abordons d'abord ses impacts en société.

2. L'industrialisation de la personnalisation

Paradoxalement, le très Big data permet de faire du très « small », à savoir calculer des profils individualisés. C'est ainsi que la campagne présidentielle d'Obama en 2012 a exploité avec succès les grandes masses de données captées à partir de sources hétérogènes et a exploré toutes les possibilités algorithmiques de personnalisation des profils des votants au point de savoir quels programmes ceux-ci regardaient et micro-cibler les messages susceptibles d'emporter leur adhésion. Le *satisfecit* émis par l'équipe *data analytics* de la campagne d'Obama témoigne d'efficacité des algorithmes :

"*Les très grandes données vous permettent d'être très fin. Elles vous permettent de faire des interventions très précises et ciblées. (...) Vous voulez que votre équipe d'analyse des données soit capable de dire aux militants : "Appelez ces numéros, frappez sur ces portes, aller dans ces quartiers." Le militant n'a pas besoin de savoir pourquoi; ils ont juste besoin de savoir qu'ils frappent sur les bonnes portes*[5]*."*

Rouvroy et Berns (2013) observent que ce phénomène conduit à une autre antinomie : une « *apparente individualisation de la statistique* » et une "*personnalisation industrielle"*. La statistique traditionnelle s'était construite sur la notion de moyenne. A partir d'un échantillon dit « représentatif », on construisait le profil moyen d'un ensemble d'observations. Cette individualisation de la statistique est une des conséquences de ce que Rouvroy et Berns (2013, 173) nomme la « gouvernementalité algorithmique » où la récolte des données massives permet en quelque sorte de 'gouverner les vivants', de conduire les affaires courantes dans divers secteurs et de prédire leurs actions futures à partir des données récoltées sur leurs actions passées. Grâce aux nombreuses traces numériques que nous laissons, chaque internaute devient son propre « profil moyen », « automatiquement attribué et évolutif et en temps réel » malgré lui, sans recours aux appareils de la statistique traditionnelle qui nécessitait un échantillonnage, des catégories socialement établies et situées

2 http://www.sdss.org/

3 http://www.galaxyzoo.org/

4 http://www.diggingintodata.org/

5 "When the Nerds Go Marching In. Accessible en ligne à http://www.theatlantic.com/technology/archive/2012/11/when-the-nerds-go-marching-in/265325/?single_page=true.

géographiquement. Cela permet à ces algorithmes de proposer à chaque internaute des formes de vie en devenir, devançant ses intentions et ses désirs en les lui livrant pré-calculés, sans qu'il y ait eu réflexivité et processus conscient de construction desdits désirs.

3. Vers une gouvernance sans gouvernés

En même temps, cette personnalisation extrême au plus près de chaque internaute se fait en s'affranchissant complètement d'une interaction explicite avec le sujet qui en est l'objet, donc sans nécessiter que l'on s'adresse directement à lui :

*« D'où un possible déclin de la réflexivité subjectivante, et l'éloignement des occasions de mise à l'épreuve des productions de « **savoir** » fondées sur le datamining et le profilage. La gouvernementalité algorithmique ne produit aucune subjectivation, elle contourne et évite les sujets humains réflexifs, elle se nourrit de données infra-individuelles insignifiantes en elles-mêmes, pour façonner des modèles de comportements ou profils supra-individuels sans jamais en appeler au sujet, sans jamais l'appeler à rendre compte par lui-même de ce qu'il est ni de ce qu'il pourrait devenir. Le moment de réflexivité, de critique, de récalcitrance, nécessaire pour qu'il y ait subjectivation semble sans cesse se compliquer ou être postposé (Rouvroy 2011)».* Rouvroy et Berns (2013, 173-4).

Ce qui intéresse la gouvernementalité algorithmique, ce n'est donc pas le sujet en « chair et en os » mais « son double statistique » ou numérique. Comme Rouvroy et Berns, (2013, 180) l'observent, il n'y a pas dé-subjectivation mais plutôt « *une raréfaction des processus et occasions de subjectivation* ». Il s'agit bien là d'un changement de paradigme car comme ces auteurs l'expliquent, l'on a une production de plus en plus algorithmique de ce qui compte comme le réel. Ce nouveau paradigme prend comme cible le potentiel, ce que les individus pourraient faire ou avoir envie de faire. On est passé d'un paradigme réel à un paradigme anticipatif. La relation entre fournisseur et client disparaît au profit d'une relation virtuelle où le sujet a l'impression d'exister alors qu'il ne fait que réaliser un scénario écrit à l'avance pour lui, en fonction de ses gestes passés. Les sujets n'éprouvant pas le sentiment d'être profilés parce que les catégories profilées ne sont ni socialement ni géographiquement déterminées, ils n'ont pas les occasions de « comparaître », de mettre en place des mécanismes pour tromper le système de surveillance ou de désobéir aux injonctions comme dans la vie réelle.

4. Quand l'ouverture des données profite au courant ultralibéral

L'ouverture des données publiques (Open Data) amplifie le phénomène du Big Data puisqu'elle contribue à la massification des données disponibles. Les discours des acteurs publics en faveur de l'ouverture des données (le G8[6] et le gouvernement Obama[7]) incitent les institutions publiques à ouvrir les données publiques sans conditions, sans préciser les modalités de cette ouverture. L'ouverture est vue comme un acte libérateur, capable à elle seule de promouvoir la démocratie et la transparence et de stimuler l'innovation. La conséquence devrait être de meilleurs services, plus de richesses et de progrès. Cette idéologie de l'ouverture comme facteur de progrès s'inscrit dans le courant altermondialiste et libertaire dont les racines puisent dans la cybernétique (Wiener, 1948). Or, l'ouverture sans une forme de contrôle, de cadrage politique et législatif peut engendrer des effets néfastes. L'ouverture des données profite essentiellement aux multinationales privées, donc au courant ultralibéral qui s'en sert pour le profilage des internautes. Ce courant est farouchement opposé à l'ouverture de leurs propres données et inventions. Les *majeurs* du web (Google, Amazon,

[6] https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex

[7] http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

Apple, Facebook, Twitter) captent tout ce qui est ouvert mais conservent en revanche jalousement, dans des formats cryptés et propriétaires, leurs données et inventions qui sont toutes brevetées. Ils communiquent agressivement contre toute tentative du courant libertaire (donc de l'ouverture) de les crawler, de les traquer en retour ou de récupérer leurs données. Les données amassées par les sociétés privées du courant ultralibéral leurs sont indispensables pour maintenir leurs positions hégémoniques, en tant que *gatekeepers* du web. Le pouvoir régalien semble avoir perdu face aux courants ultralibéral et altermondialiste qui prônent l'ouverture (Breton et Proulx, 2005). Le premier, mu par des intérêts économiques et capitalistes se nourrit de l'ouverture défendue par le second qui est plus mu par des idéaux de libre circulation et de partage des connaissances et des richesses de la planète.

5. Des traces numériques à l'identité numérique et à l'identité tout court

De cette position de *gatekeepers* du web, ces multinationales privées appliquent les algorithmes des agrégats de données à des individus, ce qui soulève des problèmes économiques, politiques, éthiques et moraux.

Les algorithmes de marketing électronique ciblé qui, sur la base des produits visionnés ou achetés par un internaute lui affichent des publicités de produits « similaires » pendant plusieurs mois, illustrent l'irruption de cette gouvernance algorithmique dans nos vies. Selon la pratique de *dynamic pricing*, en fonction de son profil numérique, calculé à son insu par les algorithmes, on peut se voir proposer un prix différent, en général à la hausse et à quelques minutes d'intervalle, pour le même voyage.

Laisser des traces numériques devient synonyme d'une normativité mais au prix d'une exposition permanente de soi. Ne pas disposer de traces numériques devient *a contrario* suspect et peut déclencher une surveillance accrue. Il n'est ainsi plus possible d'échapper à l'encerclement des dispositifs électroniques. De ces traces numériques laissées par les internautes (en fonction de ce qu'ils lisent, écoutent, regardent, achètent ou de leurs réseaux amis d'amis sur Facebook), découlent leurs identités numériques, calculées par les algorithmes, qui servent ensuite à dresser leurs « profils » qui dans certains cas, se substituent à leurs identités tout court. Si l'adage populaire « dis-moi qui sont tes amis et je te dirai qui tu es » ou le proverbe anglais « birds of the same feather flock together » nous rappellent que ce phénomène n'est pas nouveau, les algorithmes du Big Data en ont amplifié la résonance et l'impact car ils sont en mesure non seulement de construire « le profil » de chaque internaute à l'échelle planétaire, mais de les croiser de manière à produire des effets inédits et ce, à un niveau jamais atteint jusqu'ici. Les stratégies anticipatrices des agences de défense et de sécurité intérieure de nombreux pays occidentaux, qui sur la base des profils issus de ces algorithmes, peuvent arrêter et emprisonner un « terroriste potentiel » en sont les illustrations. Par ailleurs les *majeurs* du web (Google, Facebook, Twitter, Amazon) sont désormais en mesure de fabriquer des cartes d'identité numériques compte tenu des données amassées sur chaque internaute.

5. Vers des savoirs immanents aux données, sans sujets connaissants ?

Laissés en mode 'auto-pilote', les algorithmes du Big Data peuvent engendrer des effets non souhaités voire inquiétants. Les paradigmes qui ont guidé la science jusqu'au 21ème siècle nécessitaient que les scientifiques éprouvent leurs intuitions et hypothèses à l'aune d'un réel qu'ils pouvaient observer ou expérimenter. Avec le Big data et les machines dites « intelligentes », la science rentrerait dans une nouvelle ère de savoirs qui seraient immanents aux données elles-mêmes, c'est-à-dire, des savoirs qui ne seront plus produits par un processus social de questionnements, de vérifications et de contre-vérifications. Les capacités prédictives de ces algorithmes sont telles qu'ils peuvent se passer d'une question ou d'une

hypothèse initiale du chercheur et s'auto-déclencher en parcourant les données à la recherche de corrélations.

C'est là un changement fondamental dans l'approche scientifique et un renversement de paradigme. Au lieu de partir des questions et des hypothèses pour constituer des données pouvant ou non confirmer ces hypothèses comme le faisait traditionnellement la statistique, le paradigme de la fouille de données induit la démarche inverse : on part des données récoltées et on finira bien par trouver quelque chose ! Une science orientée par le *big data* peut nous livrer le « quoi » ou le « combien » sans pouvoir nous fournir les moyens de comprendre le « pourquoi » ni le « comment » des phénomènes observés. Si ce paradigme devait se généraliser, nous rentrerions dans une science sans but (intentionnalité), sans causalité (effets) et sans sujets connaissants.

Cela engendrerait selon Rouvroy (2010) « *un nouveau rapport au savoir, qui donne "l'impression d'avoir abandonné un petit peu les ambitions de la rationalité moderne, qui visait à comprendre les phénomènes en les reliant à leur cause, au profit d'une rationalité post-moderne, qui est fondée sur une logique purement statistique, donc sur la découverte de corrélations entre des données recueillies dans des contextes extrêmement divers, hétérogènes les uns aux autres, et qui sont reliés entre eux par aucun lien de causalité (...) C'est l'abandon du « savoir causal, la dévaluation de l'expérience sensible elle-même au profit du calcul*[8].»

Or, les chercheurs en statistiques et en analyse des données (*data analytics* ou *data science*) savent combien chaque algorithme de traitement de données comporte des présupposés et biais inhérents. Présentées souvent à tort comme « brutes », les données découlent en réalité d'un complexe processus de collecte, de nettoyage et de filtrage avec de nombreux biais. Jeffrey Bowker (2006) observait à juste titre que la « *donnée brute est un oxymore et une mauvaise idée et qu'au contraire, les données doivent être cuisinées avec beaucoup de soin* ». Bruno Latour (2014) proposait récemment de remplacer le terme « donnée » par « obtenue » reflétant ainsi cette réalité de « cuisine interne » des données.

6. Vers des connaissances scientifiques infalsifiables ?

Les « connaissances » produites par les algorithmes basés sur les techniques statistique et d'apprentissage apparaissent comme des connaissances objectivées car ces algorithmes apprennent constamment de leurs erreurs, affinant et perfectionnant continuellement leurs modèles, à mesure qu'ils collectent plus de données. Devant la dimension des données que seules les machines sont capables de traiter, des procédures de vérification employées habituellement par la science contemporaine deviennent inopérantes.

Or, les connaissances scientifiques ne sont pas des certitudes, elles sont conjoncturelles. Dans '*The Logic of Scientific Discovery'* (1934), le philosophe Karl Popper arguait que pour qu'une connaissance puisse être qualifiée de scientifique, elle doit pouvoir être soumise à des tests de falsifiabilité perpétuellement. Bertrand Russell affirmait que "*toute connaissance est incertaine, inexacte et partielle*[9]". Ces énoncés ne sont pas sans rappeler la théorie darwinienne de l'origine des espèces : seules des théories scientifiques les plus robustes survivent aux tests de falsifiabilité.

Si la possibilité de vérifications des conjectures scientifiques n'est plus garantie, nous serons face à un changement de paradigme important dont il faudrait s'inquiéter pour l'avenir de la science en général car on délèguerait aux machines une légitimité qui jusqu'ici était du ressort

---

[8] Questions – réponses avec Antoinette Rouvroy, *Du rôle prédictif des données à la gouvernementalité algorithmique*, 16/12/2010. http://www.internetactu.net/2010/12/16/du-role-predictif-des-donnees-a-la-gouvernementalite-algorithmique/

[9] «*All human knowledge is uncertain, inexact and partial* », Bertrand Russell, *Human Knowledge : Its scope and limits*, Simon and Shuster, New-York, (1948: 507).

des hommes, à savoir le pouvoir de nous indiquer quelles questions scientifiques méritent d'être posées sans que l'on puisse inventer des procédures pour tester la solidité de ces prédictions. Paradoxalement, alors que les algorithmes aboutiront toujours aux mêmes résultats sur les mêmes données, nous n'avons plus les moyens de soumettre les résultats aux tests de la falsifiabilité. Dans la même veine, étant basées sur des jeux de données dont on ignore les contours et les présupposés, il suffirait de varier légèrement de jeux de données ou des paramètres de l'analyse pour en varier les résultats. Les connaissances qui en découlent seraient donc doublement fragiles et conjoncturelles.

7. La fin des théories scientifiques ?

Ces algorithmes capables de se déclencher automatiquement, sans hypothèse préalable, paraissent être en mesure de livrer des résultats « justes », ayant les effets attendus. Leur usage dans le secteur du commerce électronique est à cet effet instructif. Cela soulève la question de la validité des modèles et théories scientifiques (Stedman, 2013[10]). Il est certain que de nombreux modèles scientifiques construits par des chercheurs pour représenter certaines réalités souffrent d'un sur-ajustement important (*over-fitting*), à savoir que ces modèles simplifient à outrance une réalité qui est souvent multi-dimensionnelle et complexe. La norme qui veut que toutes les expériences scientifiques soient reproductibles conduit également à une simplification extrême des conditions expérimentales en laboratoire. Cela pose des problèmes éthique et épistémologique. Il suffirait en effet de varier légèrement un seul paramètre des conditions expérimentales pour changer les résultats d'une expérience normée et donc faire tomber l'hypothèse ou la théorie qui en découle. Cela a conduit Butte (2014) à réclamer l'abandon de l'exigence de la reproductibilité des expériences scientifiques dans le domaine biomédical.

Ce constat de l'inadéquation des modèles construits par des chercheurs pour représenter une réalité a incité certains analystes à prédire que les algorithmes du Big data signeront la fin de la théorie et rendra la méthode scientifique obsolète (Chris Andersen, 2008), car des découvertes scientifiques faites en génomique n'ont pas eu besoin de modèles scientifiques qui étaient souvent faux ou inadaptés à la complexité des phénomènes observés. Andersen cite plusieurs exemples pour illustrer les performances des modèles mathématiques construits par des ingénieurs de l'entreprise Google pour profiler les internautes, sans les comprendre. Il termina son article sur une note provocatrice, se demandant s'il n'était pas temps de se demander ce que la science peut apprendre de Google. A terme, Cardon (2013, 15) craint que « *le travail d'interprétation des sciences humaines n'aurait plus de raison d'être dans l'ère des big data et pourrait être abandonné au récent mariage du behaviourisme et des algorithmes (Anderson 2008)*».

Cependant, ceux qui, tentés par le « *data fundamentalism* » et par « les illusions algorithmiques » (Boyd et Crawford, 2012) remettent en cause la pertinence des théories ou des méthodes scientifiques au profit d'une science dirigée par les données et les algorithmes, ignorent sans doute la nature profondément conjoncturelle des connaissances scientifiques. Comme l'observe Christophe Prieur (2014), «*Si l'idéal de l'apprentissage automatique est de réussir le test de Turing, c'est-à-dire qu'on n'arrive plus à faire la différence entre le comportement d'une machine et d'un humain, alors la science a perdu*».

Mais nous savons que tel n'est pas l'idéal scientifique. La science doit pouvoir comprendre ce qu'elle cherche et expliquer les méthodes avec lesquelles elle opère et comment elle aboutit à ses résultats. Avec les algorithmes du Big Data, nous perdons ce pouvoir explicatif, les algorithmes étant des suites d'opérations dont on est sûr qu'elles vont aboutir informatiquement, c'est à dire qu'elles vont converger mais à qu'il manque un modèle mathématique, à savoir

[10] Ian Stedman, Big Data and the end of the Theorist, Wired, 25 january, 2013. Accessible en ligne à http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory

que même leurs auteurs ne savent plus quelles réalités ces algorithmes modélisent.

8. *J'ai vu le 4ème paradigme et c'est nous*[11]

Dans son article intitulé « J'ai vu le 4ème paradigme et c'est nous » qui est une réponse au texte de Jim Gray, John Wilbanks (2009) rappelle que l'empirisme et la théorie sont essentiels à une bonne simulation :

« Je peux encoder une belle simulation sur mon écran dans laquelle il n'ya aucune théorie de la gravité, mais si je tente de conduire ma voiture au-delà du bord d'une falaise, l'empirisme va mordre mon derrière dans ma chute. »

Ce qui nous paraît important de souligner est que les avancées dans de très nombreux domaines sont dues à des théories, à des visions qui ont précédé les expérimentations, lesquelles, faute d'avancées technologiques, ne sont devenues possibles que des décennies plus tard. Les ordinateurs d'aujourd'hui sont des machines de Turing qui réalisent des opérations calculables que Turing avait imaginées entre 1936-1950. Alan Turing avait défini en 1936 le concept de calculabilité (*computation)* et avait posé une limite sur les opérations calculables. Il cherchait à faire reproduire par une machine théorique - car elle n'existait pas dans la réalité, ni l'ordinateur, des calculs mécaniques que le cerveau humain pouvait faire. Cela a permis de fonder la discipline de l'informatique (*computer science*) et a permis l'émergence de l'Intelligence Artificielle (IA) des décennies plus tard[12]. Ainsi, la science qui rend possible ces automatisations et ces calculs de corrélations a elle même besoin de théorie, et la machine dans laquelle les algorithmes du Big Data sont implantés est le fruit d'une théorie. L'informatique ne se réduit pas à l'ordinateur qui est certes son outil phare.

En physique de particules, la découverte du Bozon de Higgs en 2012 qui est une théorie unificatrice est d'abord la réaffirmation de la prééminence de théories scientifiques sur l'expérimentation et les corrélations des données, donc de la prééminence du schéma classique où la théorie guide la recherche et les hypothèses, même s'il faut des dispositifs expérimentaux toujours plus lourds. Nicquevert (2013) confirme que la découverte du Boson de Higgs n'aurait pas été possible sans une infrastructure technologique du Big science nécessitant une coopération mondiale et interdisciplinaire, une infrastructure technologique gigantesque, des milliers de protocoles expérimentaux, des centaines d'algorithmes et une volonté politique claire. Mais c'est bien cette théorie initiale émise cinquante ans auparavant par deux physiciens qui a motivé le projet de e-Science du *Large Hadron Collider* permettant de valider la théorie cinquante ans plus tard. Le prix Nobel de physique de 2013 a dont été attribué à ceux qui avaient imaginé l'existence de cette particule cinquante ans auparavant, sans ordinateurs et sans Big Data pour la prouver. L'expérimentation a permis d'apporter la preuve de l'existence de cette particule à 99,9999%[13].

Or, l'expérimentation toute seule n'aurait pas permis cette découverte car depuis Niels Bohr, la physique statistique a théorisé le relativisme des résultats issus de protocoles expérimentaux : à toute preuve physique est associée une probabilité. La physique a donc besoin de théorie pour guider ses expérimentations, l'expérimentation toute seule ne peut livrer la preuve irréfutable de l'existence d'un phénomène. Le Big data et ses algorithmes ne viennent pas révolutionner cela.

En mathématiques, le recours de plus en plus fréquent aux algorithmes sert non pas nécessairement à remplacer les mathématiciens mais à prouver des théorèmes anciens pour lesquels on ne disposait pas d'outils de calcul suffisamment puissants pour faire la

[11] John Wilbanks (2009)

[12] *Alan Turing,* Stanford Encyclopedia of philosophy. http://plato.stanford.edu/entries/turing/#TurMacCom. Visité le 22/04/2014e

[13] http://www.cnet.com/news/higgs-boson-theory-nets-nobel-for-pair-of-physicists/.

démonstration[14]. Là également, les données et les corrélations aveugles ne précèdent pas des découvertes mais bien le contraire, la conjecture (hypothèse) scientifique reposant entièrement sur le chercheur. Depuis 1931, le théorème d'incomplétude de Godel[15] assure qu'il existera toujours des théorèmes que la machine ne pourra pas trouver automatiquement, à supposer que nous disposions de machines à mémoire infinie.

Sur certaines problématiques, l'accumulation des données et leur traitement algorithmique peuvent bouleverser la manière de faire de la recherche et aboutir à de nouveaux éclairages. La vedette actuelle des économistes, le français Thomas Piketty, a pris le contre pied des économistes théoriciens dont il fait partie, en compilant des données sur les revenus à travers une période de 300 ans. Cela l'a conduit à remettre en cause la domination du capital sur le travail comme cause principale de manque de croissance, et à proposer la mise en place d'une taxe progressive globale sur les grandes fortunes[16]. Les données compilées sont en accès libre[17]. Bien que Thomas Piketty se réfère à des théories économiques anciennes, ce sont les données disponibles aujourd'hui dont la compilation et le traitement ont permis un nouvel éclairage et de nouvelles questions. Cependant, il a bien fallu un scientifique (en l'occurrence Piketty) pour avoir l'intuition et pour amorcer le processus de collecte des données et de recherche de tendances dans ces données. Le modèle économique proposé par Piketty est soumis au débat scientifique classique interdisciplinaire.

La science aura donc toujours besoin de théorie, qui demeure un puissant déclencheur d'investigation et de découverte scientifiques, à côté de l'expérimentation et des données empiriques.

La tendance alternative aux algorithmes, le *crowdsourcing* apparait de plus en plus déterminante pour les avancées dans des domaines confrontés au problème de Big Data mais pour lesquels les algorithmes ne parviennent pas encore à proposer des solutions acceptables à tous les problèmes qui s'y posent. La recherche en génomique fait appel de plus en plus à la foule (les humains) pour faire des annotations et la reconnaissance de motifs. Les algorithmes sont ensuite lancés sur ces annotations afin de chercher des corrélations tous *azimut*. L'enjeu n'est rien moins que le brevetage du vivant[18]. Le premier à trouver un séquençage ou un lien entre un gène et une maladie peut le breveter, d'où une course entre recherche publique et privée (courant ultralibéral), le public tentant de mobiliser très largement les citoyens bénévoles non pas pour révolutionner la génomique mais pour que les applications potentielles de ces recherches restent dans le domaine public (courant libertaire).

Le succès des méthodes qui s'appuient sur le *crowdsourcing* repose sur l'existence de « communautés » qui acceptent de produire des contenus (comme celle de Wikipedia), de valider les résultats générés automatiquement par des algorithmes (cas d'Amazon Mechanical Turk, des projets de e-science tels que Galaxy Zoo ou Telebotanica, ...) et de laisser leurs traces numériques, captées ensuite par des algorithmes qui s'en servent pour les profiler ou pour résoudre des problèmes (cas des usagers d'Amazon, de Facebook, Twitter et moteur Google).

Christophe Prieur (2014) observait que l'application de la « *data science* » en SHS devait être faite avec parcimonie car, dans les disciplines des sciences de la vie ou de l'univers, son application peut se justifier car les chercheurs de ces disciplines ne peuvent pas interroger leurs objets d'étude directement (les protéines, les cellules, les étoiles ne parlent pas !) et parfois ils ne peuvent pas s'en approcher, donc ils se contentent de les observer de loin, de les compter, des les comparer et aussi parce que, le point de vue qui importe en sciences est le point de vue

[14] http://people.math.gatech.edu/~thomas/FC/fourcolor.html
[15] http://www.math.hawaii.edu/~dale/godel/godel.html#SecondIncompleteness
[16] http://www.nytimes.com/2014/04/19/books/thomas-piketty-tours-us-for-his-new-book.html?emc=eta1
[17] http://topincomes.parisschoolofeconomics.eu/
[18] https://fr.wikipedia.org/wiki/Brevetage_du_vivant

macroscopique (climatologie, épidémiologie, ...).

En SHS, les objets d'études, ce sont nous mêmes les êtres humains, nos faits, gestes, nos dires. On peut donc s'en approcher de près pour les observer et les décrire, ce qui change la nature du problème et donc des méthodes employées. Gabriel Tarde, précurseur de la sociologie, préconisait dès début du 20ème siècle dans ses *Lois Sociales* (1903) que les SHS privilégient la vision de près. Il s'est montré critique vis-à-vis de la tendance des disciplines des sciences à privilégier des visions macroscopiques en travaillant à partir de grands volumes de données agrégées.

Après ce tour d'horizon, nous pouvons avec John Wilbanks (2009), réfuter l'idée que les données représentent le 4ème paradigme scientifique :

« *Ainsi, ce n'est pas un changement de paradigme au sens kuhnien. Les données ne sont pas en train de balayer les anciennes réalités mais elles placent une série de contraintes sur les méthodes et pratiques sociales avec lesquelles nous traitons et communiquions notre empirisme et notre théorie, et sur la robustesse de nos simulations et nos moyens d'exposer, de transmettre et d'intégrer nos connaissances. Ce qui doit changer, c'est notre paradigme de nous-mêmes en tant que scientifiques, non les anciens paradigmes de découvertes. La science orientée par les données, si elle est conduite correctement, signifiera que plus de révolutions scientifiques se produiront plus vite parce que nous pouvons confronter plus rapidement notre vision de monde contre une « réalité objective » que nous pouvons mesurer avec plus de puissance de calcul. Vu comme cela, les données ne constituent pas un « quatrième paradigme » mais une « quatrième couche du réseau », par dessus l'Ethernet, le TCP/IP et le Web.»*

On notera la conditionnalité placée par Wilbanks sur la capacité de la science orientée par les données à produire des résultats « positifs » : il faut qu'elle soit conduite « correctement » pour rendre des résultats souhaités par l'homme. Tout repose sur ce « correctement » dont l'interprétation est laissée à l'appréciation de chacun. Les potentialités de traitement des traces numériques captées par les dispositifs électroniques ne doit pas conduire au retour de ce que Cardon (2013, 17) a appelé l'« *insubmersible déterminisme technologique* ». Nous avons toujours une marge de manoeuvre mais à condition de s'emparer de la question du Big Data pour en décrypter les mécanismes et les implications.

Il ne s'agit pas de nier l'intérêt de l'automatisation de certaines tâches ou procédures routinières, ni de rejeter en bloc des méthodes algorithmiques qui produisent des résultats satisfaisants dans biens des cas (du moins des résultats dont les usagers se satisfont), mais de dire que selon les modalités opératoires dans lesquelles on se trouve, il y a des implicites et des écueils à faire remonter à la surface. On peut craindre, avec Dominique Bouiller (2014) que l'engouement actuel pour la *data science* suscitée par le phénomène du *Big Data* ne conduisent des disciplines scientifiques à ne privilégier qu'une seule modalité d'investigation, orientée par des agrégats de données massives et par la puissance de calcul des machines, les enfermant ainsi dans un paradigme scientifique unique, qui finit par rendre inaudible d'autres voies d'investigation scientifique pendant des décennies. Et l'histoire se répétera ainsi...

Bibliographie

Anderson C. (2008), « End of Theory. Will the Data deluge make the scientific method obsolete? », *Wired Magazine*: vol. 16 n° 07.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., & Messina, P. (2003), *Revolutionizing Science and Engineering through Cyberinfrastructure.* Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington, DC: National Science Foundation.

Boyd D., Crawford, K. (2012), « Critical questions for Big Data », *Information, Communication & Society*, 2012, 15:5, 662-679.

Bowker G. C. (2006), *Memory Practices in the Sciences*. Cambridge, MA: MIT Press, 273 pages.

Boullier D., Prieur C., Doueihi M., (2014), Atelier Data science, Colloque pour les 30 ans de la revue, Paris INHA, 8 au 10 janvier 2014. Texte et enregistrement de l'intervention en ligne à http://revue-reseaux.univ-paris-est.fr/fr/actualites-colloque-pour-les-30-ans-de-la-revue-reseaux/document-1775.html

Breton P., Proulx S., (2005), L'explosion de a communication : introduction aux théories et aux pratiques de la communication, Paris : La Découverte, 2005, 353 pages.

Butte A. (2014), « Translating a trillion points of data into therapies, diagnostics and new insights into disease », Keynote talk at 9th International Digital Curation Conference (IDCC 2014), San Francisco, 23-27 fév. 2014. Accessible à http://www.dcc.ac.uk/events/idcc14/video-gallery.

« Jim Gray on *eScience: A Transformed Scientific Method »* in Tony Hey, Stewart Tansley, and KrisTin Tolle (2009) (eds), *The fourth Paradigm, Data-intensive scientific discovery*, Microsoft corporation, 247 pages (pp. 19-33).

Cardon D. (2013), Présentation (pp. 9-21), in *Politique des Algorithmes. Les métriques du web*. Réseaux, fév-avril 2013, 281 pages.

Hey T., Hey J. (2006), « e-Science and Its Implications for the Library Community », *Library Hi Tech* 24(4): 515–28.

Latour B. (2014), « Le mode d'existence du Politique », Intervention orale au Colloque de IXXI "La révolution numérique et la gouvernance", ENS Lyon**,** 4 avril 2014, http://www.ixxi.fr/?p=2616 (En streaming ici : http://www.ens-lyon.eu/html/live/live.html)

Nicquevert B., (2013), Le Higgs, la chauve-souris et l'éléphant, *in* Besnier J.-M., Perriault J. (eds.) Interdisciplinarité : entre disciplines et indiscipline, *Hermès-La Revue* 67, 2013, p. 183-187.

Rouvroy A., Berns T. (2013), « Gouvernementalité algorithmique et perspectives d'émancipation : le disparate comme condition d'individuation par la relation? », *Politique des algorithmes. Les métriques du web. RESEAUX, Vol.31, n.177, pp. 163-196* (2013).

Wiener N. (1948), *Cybernetics, or the Control and Communication in the Animal and the Machine*, Cambridge, Massachussetts.

Wilbanks J. (2009), « I have seen the paradigm shift, and it is us », in *The fourth Paradigm, Data-intensive scientific discovery,* Tony Hey, Stewart Tansley, and KrisTin Tolle (eds)., Microsoft corporation, 2009, pp. 209-214.