# *Big data* - La révolution des données est en marche
## Recension de l'ouvrage

Antoine CHAMBAZ[1], Isabelle DROUET[2]
Statisticien et philosophe des sciences,
Universités Paris-Ouest Nanterre et Paris-Sorbonne

L'ouvrage dont nous proposons une recension est destiné au grand public et porte sur les données massives, ou ***Big data*** selon un anglicisme bien ancré que nous adoptons ici. Ses auteurs, Victor Mayer-Schönberger et Kenneth Cukier, sont respectivement universitaire spécialiste d'internet et des *Big data* et éditeur associé au magazine *The Economist*. Enthousiastes, ils adoptent un ton assez radical et prophétique, quoique peut-être de façon plus prononcée dans la version française. Un grand nombre de recensions antérieures sont disponibles dans des publications d'horizons divers.

Que sont les *Big data*? La variété des acceptions de l'expression est bien illustrée par la liste de quarante définitions proposées par autant de personnalités sur le blog datascience@Berkeley. La tâche de définition est d'autant plus ardue que la notion de *Big data* continue d'évoluer. S'ils ne tranchent pas définitivement la question, les auteurs proposent néanmoins des éléments de réponse. Les *Big data* peuvent être massives, mais aussi-pêle-mêle-ouvertes, exhaustives, recueillies à l'insu des sujets, désordonnées. Elles peuvent gagner à être croisées, elles sont tournées vers la prédiction. Elles ont de la valeur, et celle-ci peut se révéler bien après leur collecte. Aucune de ces caractéristiques n'est nécessaire, aucune d'elles n'est suffisante.

Toutefois, les auteurs s'intéressent moins à ce que sont les *Big data* qu'à ce qu'elles font. Les Big data permettent d'accomplir à très grande échelle des choses qui ne peuvent pas l'être à une échelle plus petite. Ce faisant, elles créent de nouvelles intuitions, de nouvelles formes de valeur, elles modifient les marchés, et les relations entre citoyens et gouvernements. Une part importante de l'ouvrage est consacrée aux transformations que les *Big data* nourrissent et au mouvement de fond de leur production effrénée, que les auteurs appellent « datafication ». Nous pourrions traduire ce terme par mise en données du réel (ou mise au pas ? Car, n'est-ce pas une vision asséchante de la réalité qui est à l'œuvre quand nous croyons capturer toute sa richesse dans une suite de 0 et de 1 ?). La construction de l'ouvrage révèle discrètement la centralité du concept de « datafication ». Il donne son titre au cinquième des dix chapitres, et ce titre est le seul néologisme quand les neuf autres sont des mots courants.

L'ouvrage est émaillé de nombreux exemples qui viennent à l'appui des thèses des auteurs. Certains sont bien connus. C'est le cas en particulier du projet *Google Translate*. Schématiquement, ce projet abandonne l'approche linguistique traditionnelle de la traduction

1. Laboratoire Modélisation aléatoire Modal'X, Université Paris Ouest Nanterre
2. Equipe de recherche SND Sciences Normes Décision, Université Paris Sorbonne

et adopte une stratégie fondée sur l'alignement phrase à phrase de textes dont nous savons qu'ils sont traductions l'un de l'autre. Le procédé est efficace grâce à la masse de traductions disponibles. Il permet notoirement de traduire d'une langue A à une langue B en utilisant de manière essentielle une langue C comme pivot. Nul besoin de disposer de traductions de A vers B ! D'autres exemples sont également remarquables. A nos yeux, l'élaboration au long cours de la géographie physique des océans entreprise par le commandant Maury, un précurseur du XIXe siècle, et la recherche raisonnée à New-York, par Flowers et son équipe, des logements sur découpés, très exposés au risque d'incendie notamment, comptent parmi les plus marquants. Le second est à double tranchant. Certes, Flowers et ses magiciens des nombres ont augmenté l'efficacité du service de prévention des incendies, et ce faisant ils ont certainement sauvé de nombreuses vies. Cependant, ils ne s'attaquent pas à la précarité du logement à la racine, c'est-à-dire à ses causes, et donc cette précarité est vouée à se perpétuer. Pour la combattre, il est impossible de faire l'économie d'une analyse qualitative, géographique et sociologique, de la ville. Certains exemples se révèlent fragiles dès le plan opérationnel, et plus particulièrement le projet *Google Flu Trends*. Si ce projet a permis de prédire l'évolution épidémique aux Etats-Unis de la grippe H1N1 plus efficacement que les statistiques gouvernementales en 2009, sa fiabilité s'est avérée décevante à l'épreuve du temps.

La thèse principale de l'ouvrage est la suivante: une révolution est en marche qui, selon le sous-titre, bouleversera la façon dont nous vivons, travaillons et pensons. Pour le dire dans les termes de Kuhn (Kuhn, 1962), nous assistons à un changement brutal de paradigme. Selon les auteurs, les caractéristiques principales de ce nouveau paradigme sont (*i*) un glissement d'intérêt du *pourquoi* vers le *quoi*, et de la causalité vers la corrélation (Chambaz et al., 2014), (*ii*) l'éventuelle exhaustivité des données, à laquelle ils associent l'expression N = *tous*, (*iii*) la possibilité désormais pleinement envisageable de laisser parler les données, comme si elles agissaient de manière autonome.

Cette analyse est discutable. Comme le souligne par exemple Strasser (Strasser, 2012), dès la Renaissance les naturalistes ont été submergés par un déluge de données et ont dû développer des techniques pour les classer et les étudier. Le XIXe siècle a lui aussi connu son *avalanche de nombres*, pour reprendre l'expression de Hacking (Hacking, 1982). Par ailleurs, il est un peu vain de comparer des quantités de données au-dessus du fossé béant qui sépare les mondes analogique et digital. En outre, la prétention à l'exhaustivité est très souvent illusoire, si ce n'est toujours, quoi qu'en disent les descriptions généralement emphatiques des bases de données massives. Leonelli démontre ainsi (Leonelli, 2014) que l'utilisation des *Big data* en biologie des organismes modèles n'a pas les caractéristiques mises en évidence par les auteurs. Elle souligne en particulier que les bases constituées sont loin de contenir toutes les données disponibles et souffrent par ailleurs nécessairement d'un biais de représentation. Finalement, indépendamment de l'exhaustivité, lorsque les masses de données sont trop conséquentes, leur manipulation requiert le concours d'artifices techniques tels que des sondages ou des projections. Nous sommes loin, à l'évidence, d'être en mesure de les laisser parler. De manière générale, il nous semble que l'analyse proposée par Mayer-Schönberger et Cukier est largement dépendante de la nature des exemples abordés, marqués par un objectif pratique plutôt qu'épistémique.

Nous aimerions maintenant revenir sur l'importance que les auteurs accordent à N = *tous*. Elle s'explique facilement par la croyance que lorsque nous avons toutes les informations relatives à un phénomène, il devient possible de faire des prédictions exactes. Ce faisant, elle témoigne du présupposé qu'il n'y a rien au-delà des données elles-mêmes, une vision de la réalité très *humienne*, dirait la philosophe, et très *pearsonienne*, dirait le statisticien. Cette vision nous semble aller à rebours du mouvement d'érosion du déterminisme, identifié par Hacking (Hacking, 1990), qui au cours des 18è et 19è siècles a peu à peu placé le hasard au centre de l'échiquier scientifique.

Selon cette vision, il n'y en particulier pas de loi immanente qui présiderait à la production des données. Par conséquent, l'objectif n'est pas de mettre au jour des traits d'une loi immanente mais seulement d'élaborer des descriptions factuelles des données disponibles et de faire des prédictions.

Pour bien mettre en lumière les ressorts de cette position, considérons l'exemple de la prédiction temporelle, c'est-à-dire de l'anticipation du futur sur la base du passé. Elle consiste à s'appuyer sur des descriptions factuelles de données recueillies à des instants successifs (nous parlons alors de données temporelles), puis à déterminer les tendances qui s'en dégagent et, enfin, à les prolonger. Dans la théorie statistique de l'estimation séquentielle, c'est justement l'existence d'une loi immanente qui justifie l'opération de prédiction, éventuellement au prix d'hypothèses dont la formulation même engage une référence à cette loi. Dans cette théorie, différents prédicteurs peuvent être mis en concurrence par comparaison des risques qui leur sont associés, le risque d'un prédicteur étant l'écart moyen, sous la loi immanente, entre la valeur à prédire et sa prédiction. Parce qu'elle suggère l'absence de loi immanente, l'importance accordée par les auteurs à N = *tous* les met ainsi en porte-à-faux avec la théorie statistique de l'estimation. Elle les rapproche, en revanche, de la théorie de la prédiction de suites individuelles, où les observations sont le produit d'un certain mécanisme inconnu, non spécifié, pouvant être tout autant déterministe que stochastique, ou même dynamique et antagoniste. Dans ce cadre, ce n'est pas le risque qui est la mesure centrale de qualité d'un prédicteur mais le regret, qui consiste en une comparaison entre les gains effectifs générés par le prédicteur et ce qu'auraient pu être les gains si le prédicteur optimal avait été choisi. Notons qu'ici, le hasard est réintroduit de la main des prédicteurs eux-mêmes: chassez le hasard, il revient au galop ! La randomisation se révèle en effet essentielle pour déterminer un prédicteur optimal.

Bien qu'ils soient partisans de laisser parler les données, les auteurs nous préviennent: "nous devons nous garder de trop nous reposer sur les données pour ne pas répéter l'erreur d'Icare qui, subjugué par la prouesse technique du vol, ne l'a pas maîtrisé et s'est abîmé en mer". Loin des considérations épistémiques qui ont été les nôtres, ils situent les limites du recours aux *Big data* sur le seul terrain réglementaire ou légal. Les inquiétudes qu'ils expriment concernent principalement la vie privée et les libertés individuelles. Un mouvement de fond économique et sociétal est assurément en marche. Il nous appartient de le contrôler et, surtout, de nous l'approprier. L'État français a ainsi créé une plateforme ouverte des données publiques françaises[3] et appelle à y contribuer librement. Nous faisons collectivement face à un enjeu fondamental d'éducation et de sensibilisation de tous les citoyens, jeunes et moins jeunes. Ensemble, domptons les *Big data*, ne nous laissons pas dévorer par elles. Nous concluons la recension sur cette note résolument optimiste.

## Références


Chambaz, I. Drouet, and J-C. Thalabard. *Causality, a trialogue*. Journal of Causal Inference, 2 (2): 201-241, 2014.

Hacking. *Biopower and the avalanche of printed numbers*. Humanities in Society, 5 (3-4): 279-295, 1982.

Hacking. *The taming of chance*, volume 17. Cambridge University Press, Cambridge, 1990.

T. S. Kuhn. La structure des révolutions scientifiques. 1962.

S. Leonelli. *What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society*, 1 (1): 1-11, 2014.

J. Strasser. *Data-driven sciences: From wonder cabinets to electronic databases. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43 (1): 85-87, 2012.


3. https://www.data.gouv.fr/fr/