# La recherche scientifique à l'ère des Big Data
## Cinq façons dont les Big Data nuisent à la science et comment la sauver

## de Sabina LEONELLI (2019)

Antoine ROLLAND[1]
Université Lumière Lyon 2


**Livre** (118 pages – Traduit de l'italien)
**Auteurs :** Sabina LEONELLI
**Édition :** Mimesis (Collection : Philosophie) – 2019
**ISBN :** 978-8869761843

Le court livre de Sabina Leonelli aborde la question de l'influence des Big Data dans la pratique scientifique actuelle. Il est constitué de quatre chapitres définissant les Big Data, donnant à voir la manière dont ces dernières nuisent à la recherche scientifique, proposant des solutions rapides à mettre en place et indiquant des pistes de travail à long terme pour une science participative et responsable.

Sabina Leonelli est philosophe des sciences et a beaucoup travaillé avec des biologistes. Elle s'intéresse aux systèmes créés au cours de l'histoire pour concevoir des descriptions et des explications du fonctionnement du monde. Elle a donc une approche philosophique et épistémologique de la manière dont les données impactent la recherche scientifique.

Le chapitre 1, intitulé « Que sont les Big Data ? », en propose une définition. Partant des « V » classiques quand on parle du domaine (ici, sept « V » : Variété, Volume, Vitesse, Véracité, Valeur, Volatilité, Validité), elle présente les « Big Data » comme des données de types et de provenances différents, mises en relation les unes avec les autres, et qui permettent d'instaurer

1. antoine.rolland@univ-lyon2.fr

des connexions entre des secteurs et des approches qui ne dialoguaient pas habituellement. Pour l'auteure, la disponibilité de quantités phénoménales de données fait que la science est en train de passer d'une approche centrée sur la théorie à une approche centrée sur les données. Les données peuvent potentiellement être porteuses de connaissance indépendamment de leur rôle de preuve pour une hypothèse théorique déterminée.

Dans cette perspective, Sabina Leonelli pointe dans le chapitre 2 intitulé « Signaux d'alarme » cinq dangers pour la science à ne fonctionner qu'à partir de données :

- le conservatisme, et le fait de n'utiliser que des données anciennes ;
- le manque de fiabilité, et l'utilisation de données douteuses, de mauvaise qualité ;
- la mystification, ou le problème de n'utiliser qu'une partie des données disponibles, typiquement celles venant renforcer une hypothèse *a priori* ;
- la corruption, le fait de mettre volontairement des données fausses ou de mauvaise qualité à disposition. A cet égard le risque est grand que la commercialisation des données s'accompagne d'une séparation croissante des données elles-mêmes de leur contexte, sans aucune possibilité de recontextualisation ;
- le problème des données sensibles, non seulement d'un point de vue individuel, mais aussi d'un point de vue d'un groupe de citoyens, ce qui élargit notablement l'ensemble des données définies comme sensibles.

Pour contrer ces dangers, Sabina Leonelli développe dans le chapitre 3, intitulé « Comment éviter le pire », une approche relationnelle pour l'épistémologie des données. C'est là le cœur de l'approche originale de l'auteure. Dans la vision « représentative » classique, les données sont des représentations fiables de la réalité, produites par des interactions entre l'homme et le monde : elles sont la porte d'entrée pour accéder au monde de manière systématique, contestable et reproductible par d'autres, contrairement à la connaissance empirique du monde qui est tirée de notre perception sensorielle du monde. Du point de vue de l'auteure, l'approche représentative voit les données comme des faits incontestables et privés d'aspects théoriques et subjectifs, sans tenir compte de l'histoire des données, de leur provenance, des circonstances conceptuelles, matérielles et sociales dans lesquelles on peut les interpréter. Au contraire, dans une approche « relationnelle », les données sont conçues comme des objets mis en relation avec une question irrésolue, pour des raisons qui dépendent de la situation dans laquelle la question est posée, et non pas comme une représentation du réel. Dans cette vision relationnelle des données, l'histoire des données, la manière dont elles ont été construites, stockées, transmises au fil du temps, et la manière dont on peut les percevoir comme aidant à trouver une solution à un problème posé dans un contexte donné sont aussi importants que la donnée en elle-même. Le processus d'obtention de connaissance inclut alors des objets choisis pour remplir la fonction de données, mais en relation avec d'autres éléments cruciaux pour l'interprétation, comme l'objectif de la recherche, les hypothèses conceptuelles…

Enfin, dans le chapitre 4, l'auteure développe des pistes pour éviter de se résigner au déterminisme technologique. Elle plaide pour l'introduction de l'éthique dans la recherche scientifique, reconnaissant au chercheur une responsabilité dans l'usage (ou le mésusage) qui est fait du résultat de ses recherches. Elle plaide aussi pour l'importance de ralentir les temps de recherche, ainsi que l'ouverture au dialogue et à la confrontation sociale la plus large possible pour l'usage des données massives à des fins de recherche. La slow science constitue une alternative au modèle d'utilisation des données soumis à une pression constante de l'environnement de la recherche.

Sabina Leonelli suggère pour finir huit principes pour faciliter la transformation des données massives en connaissance fiable. Elle reprend dans ces huit principes les thèmes développés précédemment : la nécessité d'avoir des données de qualité et des compétences en gestion de données, d'avoir une liberté de recherche et l'accès à des sources de données aussi diverses

que possible, et de conduire les recherches scientifiques de manière éthique et en dialogue avec leur environnement. En un mot, il s'agit de bien considérer les données de manière relationnelle, en interrogeant autant que possible leur provenance et leur histoire.

## Mon avis

Voilà un petit livre intéressant qui cerne la question de l'utilisation intensive des « Big Data » dans le monde de la recherche. S'il peut faire ouvrir les yeux au grand public, ou à des chercheurs peu familiers du traitement de données, il ne surprendra pas le statisticien dans son constat, ses conclusions et ses recommandations. Au contraire, le praticien statistique se verra renforcé dans ses bonnes pratiques : ne pas considérer que les données sont « données » justement, mais toujours interroger leur provenance, leur qualité, leurs biais possibles, leur capacité à répondre à la question posée… L'approche relationnelle prônée par Sabina Leonelli met des mots sur ces « bonnes pratiques » statistiques, qui ne concernent finalement pas seulement les données massives mais tout travail sur des données. Il faut maintenant espérer que les décideurs et financeurs de la recherche lisent ce livre, et ainsi cessent d'être tentés par l'idée selon laquelle les données massives seraient la solution gratuite et miraculeuse à tous les problèmes.