



Fiche de lecture

Big, Fast, Open Data. Sous la direction de Yannick Lejeune, EPITA Edition, Paris, 191 pages, octobre 2014.

Fiche rédigée par Daniel Ventre. 3 Décembre 2014

Sous la direction de Yannick Lejeune, 23 auteurs traitent de la notion de « données » dans un ouvrage structuré en 5 chapitres :

- L'humain et ses données, le « quantified self »
- L'informatique des données et les mondes numériques
- Société et gouvernance : le citoyen et la cité
- Les entreprises et le business des données
- La science à l'ère des mégadonnées

Nous retenons de cette lecture quelques arguments qui nous paraissent refléter l'essentiel des principaux messages véhiculés au travers du livre.

Sur les modifications, conséquences, voire révolutions induites par la data (big, fast, open)

- Les rapports qu'entretient l'homme avec ses données sont modifiés. Les big data, fast et open data, transforment le rapport à soi.
- Le big data transforme notre manière de vivre et de penser. Il révolutionne notre rapport au monde : on ne collecte plus ni ne traite de petits échantillons, des sous-ensembles de données. Le changement d'échelle entraîne nécessairement un changement de point de vue.
- Aujourd'hui tout autour de nous produit des données. Ces données se réfèrent aux comportements humains. Les données sont partout : directes, ouvertes, prédictives... Explosion du volume de données produites car explosion des relations individus-systèmes, du nombre de capteurs, senseurs. Cette quantité de données permet de voir le monde sous

un angle nouveau : on observe le monde au travers des données. Mais davantage que la masse de données, c'est la complexité qui définit le big data (p.151). Le big data se définit grâce aux trois mots : analyser, prédire, réagir (p.137).

- Le big data montre des corrélations (connexions apparentes entre des données) et non une relation de causalité
- La capacité à collecter, traiter, produire des masses énormes de données nous libère des contraintes existant jusqu'alors. On peut désormais s'appuyer sur des quantités de données bien plus massives. Mais quel que soit le volume de données, celles-ci ne sont jamais le reflet de l'entière réalité, elles sont toujours incomplètes, imparfaites (p.20)
- Les problèmes, enjeux, défis liés au Big Data sont multiples : surveillance (p.51) ; risque d'abus dans l'utilisation des approche probabilistes, prédictives ; risque de dictature de la donnée. Doug Laney¹ a défini les problématiques propres à ces données sous le célèbre principe des 3V : volume, vitesse, variété. D'autres attributs peuvent être ajoutés : véracité, variabilité, valeur...
- Importance de l'accessibilité à la donnée (pas de big data lorsque l'accès est limité, contraint)
- Le problème n'est pas tant la production et la collecte des données, que leur manipulation et leur interprétation, leur utilisation.

Le point de vue du juriste

Le droit à la propriété des données n'existe pas aujourd'hui dans le monde (p.105). Or si elles n'appartiennent à personne, il ne peut pas y avoir vol. Il faut créer le droit à la souveraineté, à l'autodétermination par l'utilisateur de ses droits (p.106), c'est-à-dire droit de pouvoir contrôler ses données et décider de leur utilisation.

Les défis posés par le big data du point de vue juridique sont principalement les suivants: comment appliquer la réglementation informatique et liberté ; comment faire respecter le droit à l'oubli (tout en respectant le droit à l'histoire) (p.109). Le droit des algorithmes reste à inventer (p.110).

Le point de la vue de la Défense

Ce qui intéresse l'auteur, c'est la dimension humaine du big data : la donnée personnelle, qui explose. « La mission fondamentale de l'Etat est de protéger les citoyens, ici les données des citoyens » (p.87). Il distingue 3 cercles de souveraineté : sur la donnée personnelle, des entreprises/organisations, des Etats. Globalement le chapitre dédié au point de vue de la Défense ne traite guère du big data.

D'autres parties du livre évoquent (survolent) les usages du big data à des fins de sécurité (p.51 et suiv.) : le big data contribue à la surveillance, au « cyberint » (cyber intelligence) c'est-à-dire au renseignement de masse. L'un de ses méthodes consiste à collecter massivement des données, sur une longue période, et à mesurer les différences, puis tirer des conclusions, déclencher des signaux d'alerte, dès qu'une telle différence est identifiée.

Sur la définition des concepts

Ce livre est l'occasion, pour tous ceux qui sont peu familiers de ces sujets, de découvrir quantité de concepts : bio-informatique, dataïsé, homo-data-sapiens, algorithmiste, soi quantifié (quantifier tout ce qui se passe à propos de soi-même), médecin data-scientist ... La nouveauté (la révolution de l'humanité en cours ?) se dit visiblement mieux en anglais: big data, open data, fast data, open

¹ Doug Laney, 3D Data Management : Controlling Data Volume, Velocity and Variety, Meta Group, 6 février 2001, 4 pages [<http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>]

access, data scientist, quantified self, quantified others (données qui portent sur les autres), sport-scientist, dispractices (mauvaises pratiques), frames (images par seconde), open government, etc.

- Open data : données ouvertes par les administrations dans un premier temps. Ce processus n'est pas lisse, homogène, n'est pas naturel. Il y a de nombreuses résistances.
- Open access : libre accès aux publications scientifiques
- Open science : open access + open data + logiciel open source + recherche participative et contributive.
- Open web : considérer les connaissances comme des biens communs
- Les fast data : celles qui arrivent en temps réel. Peu de lignes sont accordées aux fast data elles-mêmes dans cet ouvrage, qui se concentre en réalité sur le big data et l'open data.

Sur la « valeur » de la donnée

Il est à maintes reprises dans l'ouvrage question de la « valeur » de la donnée :

- elle ne réside plus seulement dans l'objectif pour lequel elle a été collectée, mais dans les utilisations et réutilisations possibles ultérieurement (p.19)
- la valeur des données réside avant tout dans la capacité à les utiliser intelligemment (p.182)
- la notion de valeur a priori des données (n'a pas) beaucoup de sens. C'est la contextualisation qui confère de la valeur aux données (p.141).

Commentaires

Les regards portés sur les transformations induites par l'explosion des données hésitent entre un solutionnisme technologique, non pas simplement webcentré tel que le décrit et critique Evgeny Morozov par exemple, mais mathématico-centré, plus précisément centré sur l'algorithme. Les algorithmes (re)deviennent centraux. Ils permettent de faire parler les données et d'en produire de nouvelles. La généralisation du big data fait émerger de nouveaux algorithmes (p.151). Exit les outils de gestion de bases de données, traditionnels (les auteurs ne vont pas jusqu'à les qualifier d'archaïques). On s'interroge sur le pouvoir qui leur est conféré : « s'oriente-t-on vers la gouvernance algorithmique ? » (p.185)

Vision technocentrée utopique : le bonheur est à portée de main grâce à cette nouvelle évolution technologique (le même discours était tenu aux balbutiements de l'internet, des autoroutes de l'information, qui devaient rendre l'humanité prospère). Grâce aux données, aux mathématiques, « nous sommes aptes à découvrir le fonctionnement réel de notre société » (p.22). Les maîtres de ces données et technologies (les data scientists) seraient donc les nouveaux maîtres du progrès, si ce n'est de l'humanité ? Le big data apparaît comme une solution à bien des problèmes : avec les objets connectés, la société deviendra intelligente (smart cities, smart cars, smart phones, ...), et le big data permettra « d'améliorer l'organisation du pays... rendre le système de télécommunications bien plus efficace... le système de santé bien plus robuste... améliorer les services de transport » (p.26). Le big data peut améliorer nos sociétés (p.26). La data redéfinit les rapports de force dans le monde commercial, industriel. Le big data est l'avenir du marketing (p.143). Le bonheur, le progrès, passent par les données : « tout peut être dataisé » (p.27), l'ouverture des données c'est la démocratie (p.75). L'un des auteurs rêve d'une société qui sera plus quantitative (p.29). Le bonheur et le progrès

par la science, et surtout par les chiffres, les mathématiques. « Nous allons construire un monde meilleur » (p.36). On a déjà entendu cela... On y lit même que les sciences humaines, de « bonne science », c'est-à-dire essentiellement qualitatives, deviendraient véritable science, parce qu'elles vont désormais utiliser les données massivement, se servir du big data, devenir réellement quantitative (p.22) Les sciences humaines seront plus précises et plus prédictives (p.29), elles « deviennent une vraie science ».

L'ouvrage se termine sur des considérations plus humbles, plus retenues. Le big data ne fait pas tout. Il produit certes de la donnée, mais « un même savoir produit des effets bien différents » (p.188), accordant encore à l'être humain le choix, le pouvoir de décision. Car les algorithmes ne sont pas des entités autonomes : les résultats qu'ils produisent sont aussi le reflet des stratégies qui ont gouverné leur conception.

Chaire Cyber-Défense et Cyber-sécurité

Fondation Saint-Cyr, Ecole militaire, 1 place Joffre, 75007 Paris
Téléphone: 01-45-55-43-56 - courriel: contact@chaire-cyber.fr; SIRET N° 497 802 645 000 18
La chaire remercie ses partenaires



CENTRE DE RECHERCHE
des ÉCOLES DE
SAINT-CYR COÛTQUIDAN



THALES