

Section 1 : qu'est-ce qu'une donnée ? Petite histoire sociale des données et de leur exploitation

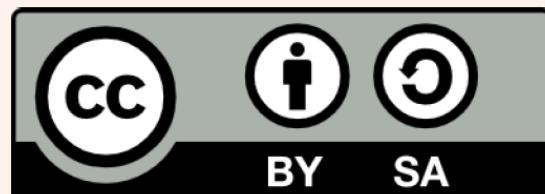
Culture générale des données

Dataactivist, 2018-2019

Ces slides en ligne : <http://dataactivist.coop/SPoSGL/section1.html>

Sources : <https://github.com/dataactivist/SPoSGL/>

Les productions de Dataactivist sont librement réutilisables selon les termes de la licence [Creative Commons 4.0 BY-SA](#).



Plan du cours

1- "Au fait, c'est quoi une donnée ?"

Ecoutez "l'interview de Serge Abiteboul, commissaire scientifique de l'exposition Terradata et directeur de recherche à l'Inria"

2- L'industrialisation de la production des données

Regardez l'interview des créateurs de Visicalc, le premier tableur

3- La fin de la science ?

Lire l'article " Big Data : est-ce que le déluge de données va rendre la méthode scientifique obsolète ? "

Quizz section 1

1 - Au fait, c'est quoi une donnée ?

Introduction

Ecoutez l'interview de Serge Abiteboul, commissaire scientifique de l'exposition Terradata et directeur de recherche à l'Inria



On a toujours voulu inscrire,
retenir des données,

Tous droits réservés

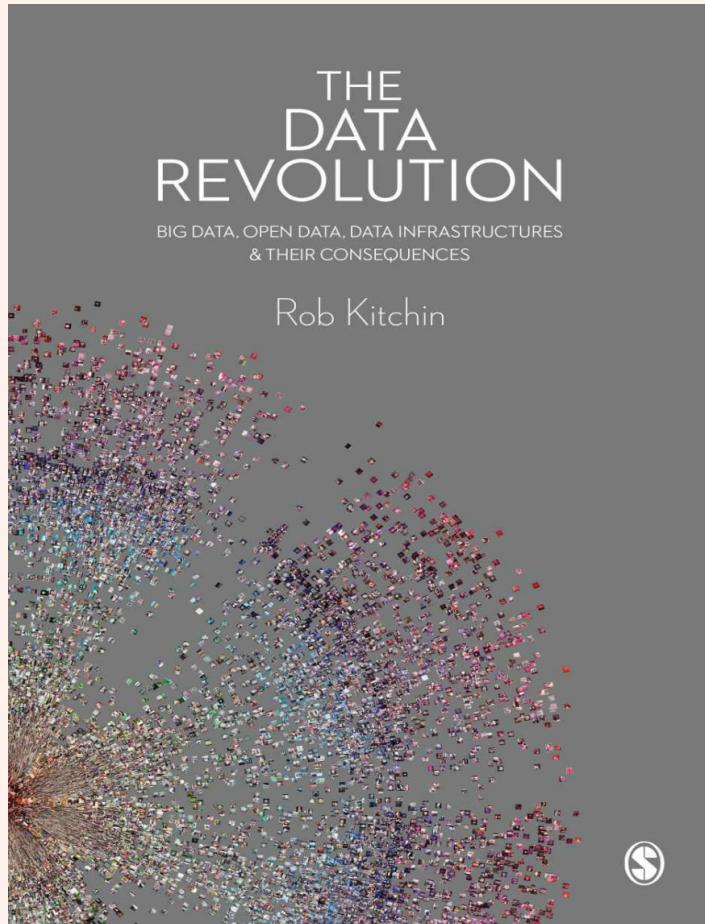
Les données sont partout !



Identifiez autour de vous 3 appareils qui collectent des données

source : The Economist

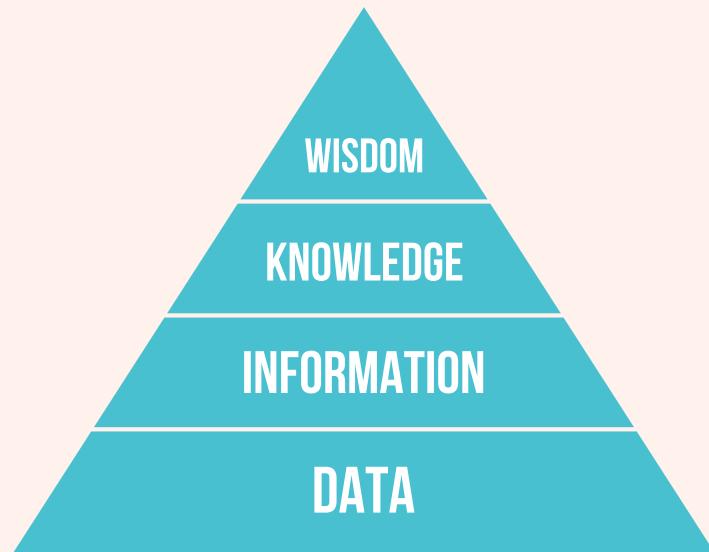
Une définition des données



Les données sont couramment comprises comme les matériaux bruts produits dans l'abstraction du monde en catégories, mesures et toute autre forme de représentation-nombres, caractères, symboles, images, sons, ondes électromagnétiques, bits qui constituent les fondations sur lesquelles l'information et le savoir sont créés.

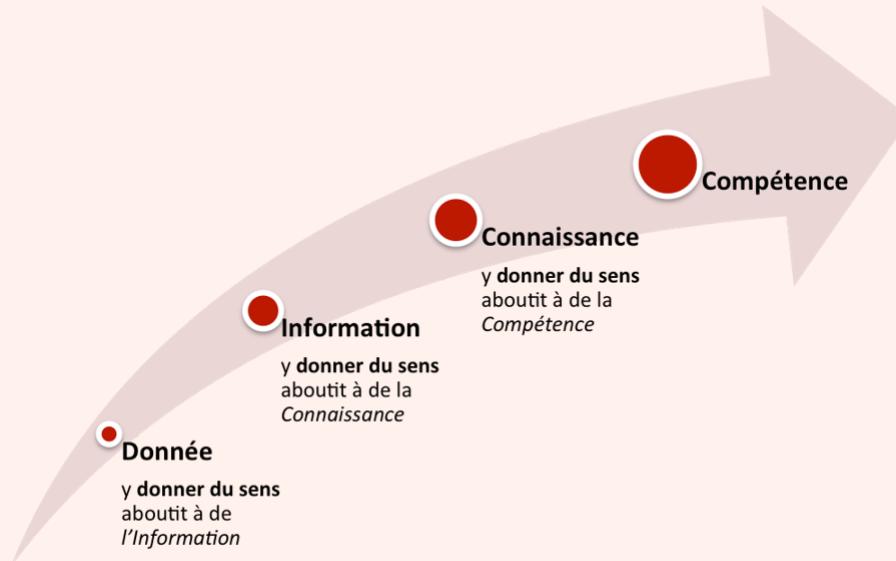
La pyramide Data-Information-Knowledge-Wisdom

Attribuée à [Russell Ackoff](#) en 1989, elle signifie que :



- Les **données** sont la matière "brute" de l'information conçues plutôt pour des machines.
- **L'information** pourrait être définie comme des données qui ont été interprétées pour dégager du sens pour des humains.
- En donnant du sens à de l'information, on obtient de la **connaissance**
- En donnant du sens à la connaissance on obtient de la **sagesse**.

La pyramide Data-Information-Knowledge-Wisdom



NB : le haut de la pyramide, est parfois remplacé par "compétence"

Les données, c'est aussi tout ce qui circule dans un ordinateur

Les données ne sont pas seulement le fondement du savoir, elles sont aussi la base de l'informatique. Tout ce qui circule dans un ordinateur, ce sont des données.

Data inflation		
Unit	Size	What it means
Bit (b)	1 or 0	Short for “binary digit”, after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10} bytes	From “thousand” in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From “large” in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From “giant” in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From “monster” in Greek. All the catalogued books in America’s Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America’s postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of <i>The Economist</i>
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures.
Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: *The Economist*

Pensez à votre abonnement téléphonique, chaque mois, vous payez pour consommer un certain volume de données quantifié en octet ou en bit.

Le volume des données créées et traitées ne cesse de croître en même temps que les capacités de calcul des ordinateurs.

2 - L'industrialisation de la production des données

La tablette mésopotanienne : inscrire une réalité complexe

Vers 3200 av. J.-C., en Mésopotamie, la civilisation sumérienne a inventé l'écriture d'abord pour mémoriser des comptes (difficile de recenser des têtes de bétail ou des sacs de grains oralement).

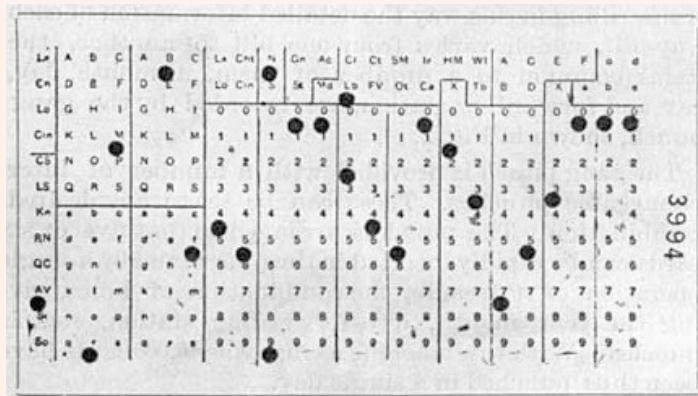


"La compatibilité a été l'une des premières *success story* de l'écriture ; les premières tablettes comprennent souvent des listes de compte."

Source : Abiteboul & Peugeot (2017). *Terra Data : qu'allons nous faire des données numériques ?*, Paris : Le Pommier.

La carte perforée (1884) : le début de la massification des données

Apparue au départ dans les métiers à tisser, les carte perforées contiennent des informations représentées par la présence ou l'absence de trous dans une position donnée.

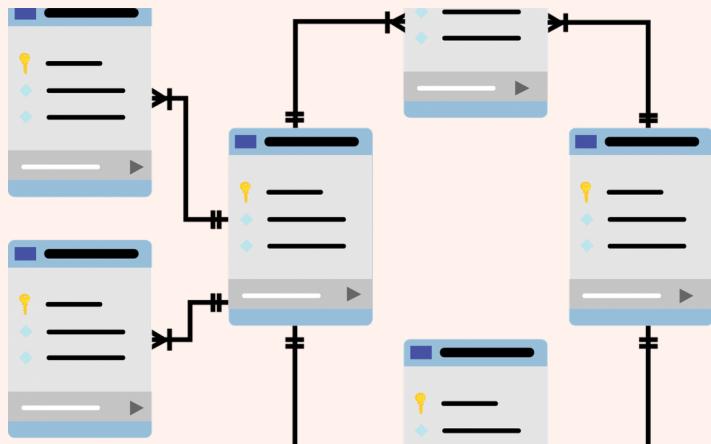


Elles sont les premières mémoires de masse utilisées dans l'informatique au XIXe siècle.

En 1884, Herman Hollerith a déposé un brevet pour une machine à cartes perforées destinée à accélérer la production de statistiques pour les gouvernements. Deux ans plus tard, il crée IBM le géant de l'informatique.

Les bases de données relationnelles (1970)

Les bases de données relationnelles facilitent grandement le traitement des données puisqu'elles paraissent à travers une interface utilisateur : « il faut protéger les futurs usagers de grandes banques de données d'avoir à connaître comment les données sont organisées dans la machine » (Codd 1970).



D'un point de vue physique, les données sont inscrites dans des tables et reliées entre elles par un schéma et des identifiants uniques. Cela permet de traiter de plus grands volumes, de développer des données plus complexes et d'éviter des erreurs de saisie.

Le tableur (1979) : *data to the people*

En 1979, Dan Bricklin, un ancien analyste financier exaspéré par les techniques de calcul encore manuelles, a imaginé une technique de calcul visible (« *Visible Calculator* »).

Son logiciel "Visicalc", **démocratise la production des données** en proposant le système de la feuille de calcul sur laquelle les données peuvent être directement manipulées :

"La facilité d'utilisation de Visicalc provenait du fait que l'utilisateur n'avait pas besoin de connaître de langage de programmation. Sur cet aspect, Visicalc était **l'équivalent du traitement de texte** dans lequel un utilisateur arrange directement l'impression de la page, à l'opposé des systèmes d'écriture où l'utilisateur devait inscrire un ensemble d'inscriptions pour mettre en page le texte." ([Campbell-Kelly, 2007](#))

Le tableur (1979) : *data to the people*

Dan Bricklin et Bob Frankston ont inventé le tableur, le premier logiciel de calcul de masse. Découvrez en 5 minutes leur invention :

Before Excel there was VisiCalc: An interview with its creat...



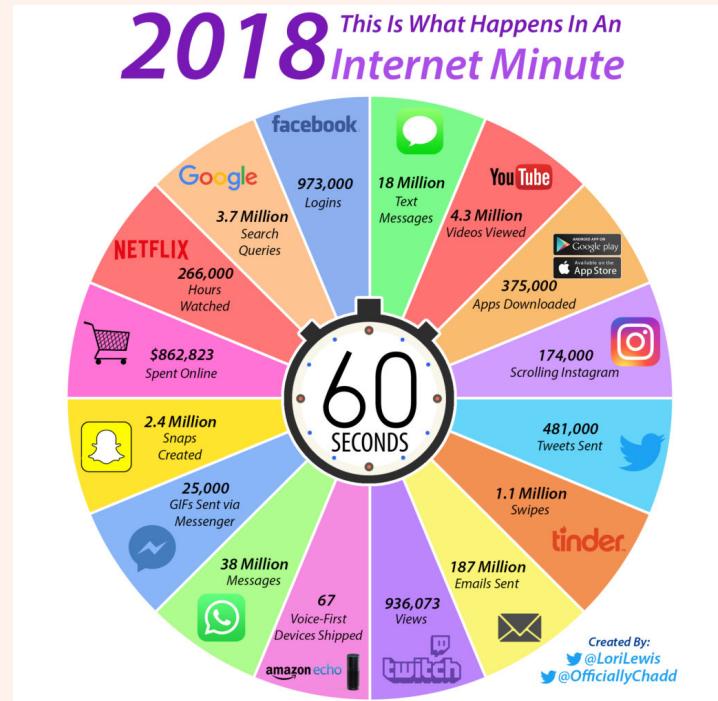
3 - la fin des sciences ?

Le déluge des données

La réflexion autour de la fin de la science part du constat de l'explosion de la production de données comme l'illustre cette infographie sur les réseaux sociaux en une minute.

"Avec suffisamment de données, les chiffres parlent d'eux-mêmes."
Chris Anderson,
journaliste *Wired Magazine*

Source



C'est nouveau ce déluge ?

"Les perceptions d'une "surabondance informationnelle" (ou d'un "déluge de données") ont émergé de manière répétée depuis la Renaissance jusqu'aux périodes modernes et, à chaque fois, des technologies spécifiques ont été inventées pour gérer la surabondance perçue."

Strasser, B. J. (2012). "Data-driven sciences: From wonder cabinets to electronic databases""

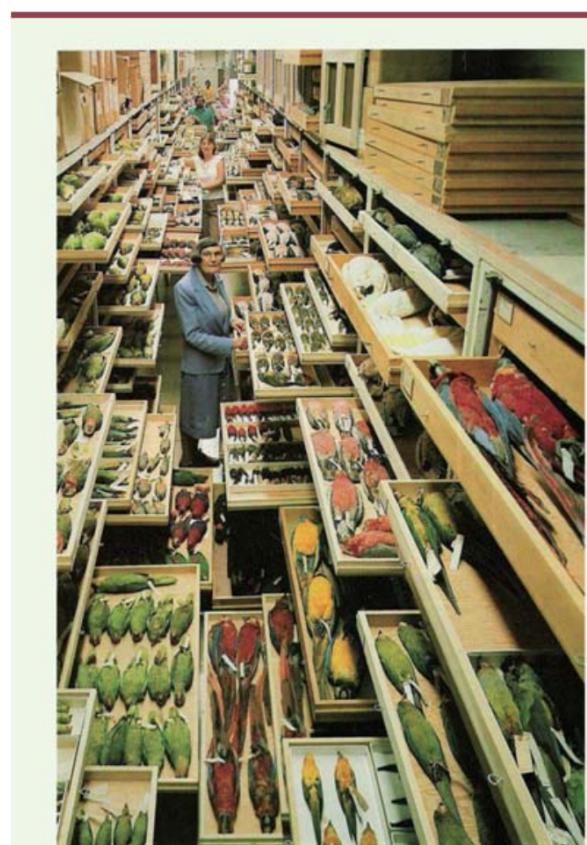


Figure 5. Collection ornithologique du Smithsonian National Museum of Natural History (© Chip Clark, SI Photo Services).

La méthode scientifique est-elle obsolète ?

Les sciences sont traversées par la promesse d'un **quatrième paradigme scientifique** qui remplacerait le modèle hypothético-déductif par l'analyse de données. Il suffirait alors d'**explorer les données pour identifier des corrélations** (une relation entre des phénomènes) et de **comprendre la causalité**.

Le déluge des données rend la méthode scientifique obsolète, l'analyse des motifs et des relations contenues dans les données massives produit intrinsèquement un savoir significatif et éclairé sur des phénomènes complexes. Il y a maintenant une meilleure manière de faire. Les petabytes nous permettent de dire que « la corrélation suffit ». Nous pouvons analyser les données sans hypothèses sur ce qu'elles peuvent montrer.

Anderson, C. (2008) "The end of theory: The data deluge makes the scientific method obsolete", *Wired*

Le risque : confondre corrélation et causalité

Deux événements (appelons les X et Y) sont corrélés si l'on observe une relation entre les deux. Une erreur de raisonnement courante consiste à dire : « X et Y sont corrélés, donc X cause Y ». On **confond corrélation et causalité** car en réalité, il se pourrait aussi que Y cause X, ou bien que X et Y aient une cause commune Z, ou encore que X et Y soient accidentellement liés mais n'aient aucun lien de causalité.

L'effet cigogne désigne la tendance à confondre corrélation et causalité.

"Dans les communes qui abritent des cigognes, le taux de natalité est plus élevé que dans l'ensemble du pays. Conclusion : les cigognes apportent les bébés ! En fait, les cigognes nichent de préférence dans les villages où la natalité est plus forte en milieu rural que dans les villes."

Source : cortecs.org

Vous aussi, générez des corrélations absurdes

L'équipe des Décodeurs du *Monde* a produit un **générateur de comparaisons absurdes et parfois drôles**, essayez le !

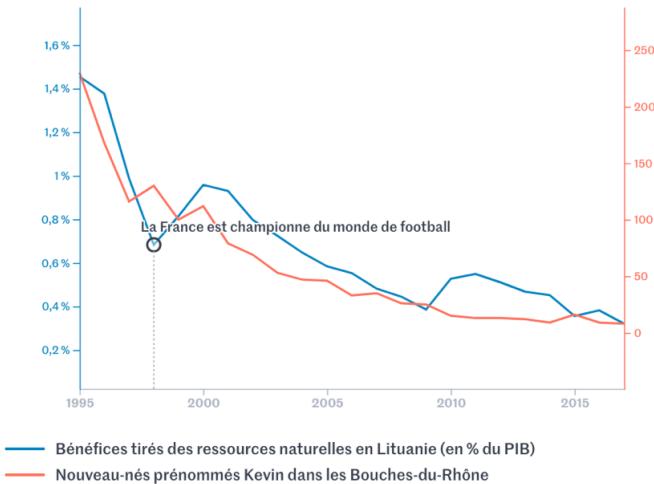
LES DÉCODEURS • DATAVISUALISATION

Corrélation ou causalité ? Brillez en société avec notre générateur aléatoire de comparaisons absurdes

Internet voit parfois émerger des courbes ou des cartes qui prétendent pouvoir expliquer simplement des questions complexes : quelques conseils pour ne pas tomber dans le panneau.

Par Pierre Breteau, Maxime Ferrer et Lucas Baudin • Publié le 02 janvier 2019 à 07h11 - Mis à jour le 06 mars 2019 à 10h28

Cliquez sur le bouton pour générer une nouvelle comparaison au hasard : ici, nous avons un coefficient de corrélation de 93,5 % entre les deux courbes.



Le déluge des données à l'épreuve des sciences sociales

Les sciences sociales n'échappent pas à la montée en puissance de la science *data-driven*. Le *social computing* désigne une branche de l'informatique qui essaie de comprendre les comportements sociaux par l'analyse de données et l'usage d'outils informatiques.

Lev Manovich (2011) signale que **cette approche comporte plusieurs risques** :

- elle favorise les chercheurs ayant des liens officiels avec les industriels des réseaux sociaux qui vont fournir les données (difficile alors de les critiquer) ;
- des évidences pour les sciences humaines vont être présentées comme nouvelles ;
- à l'inverse, certains enseignements majeurs de la littérature des sciences humaines sont ignorés ;
- les traces numériques des réseaux sociaux sont perçues comme authentiques ignorant les multiples stratégies de gestion des identités des individus ;
- ces recherches disposent d'une force rhétorique bien supérieure en s'appuyant sur les données de plusieurs millions d'individus.

Merci !

Contact : samuel@dataactivist.coop