UCLouvain

FIAL
Un éventail de formations, une multitude d'horizons

Big Data

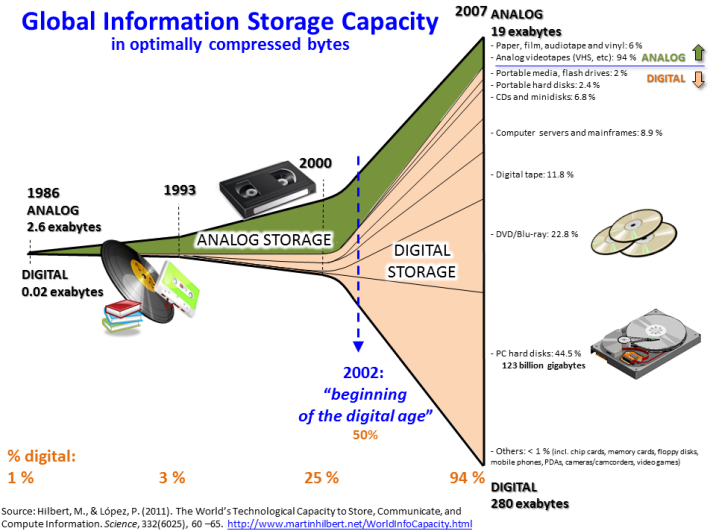
LSC1120A
séance 3

Moteurs du changement

Moteurs technologiques :

- L'espace mobile
  - 4G, bientôt 5G
  - Tracking de localisation omniprésent
  - Utilisation constante d'apps

Capacité de stockage mondiale

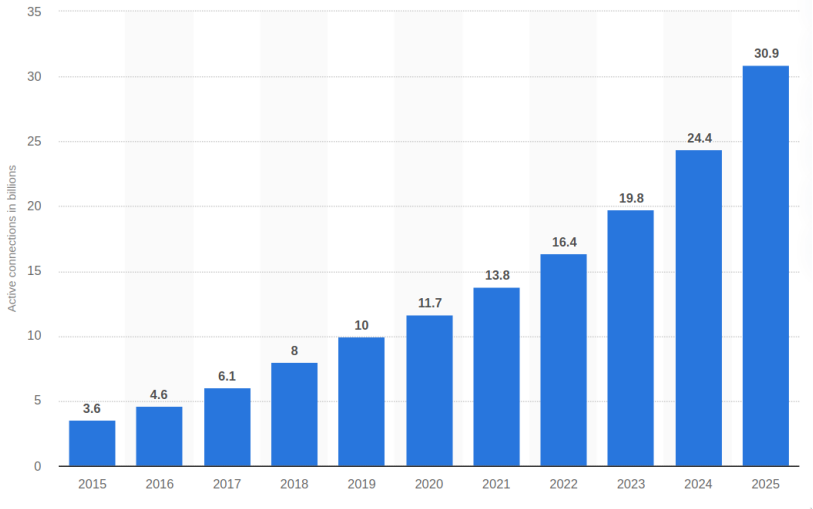


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

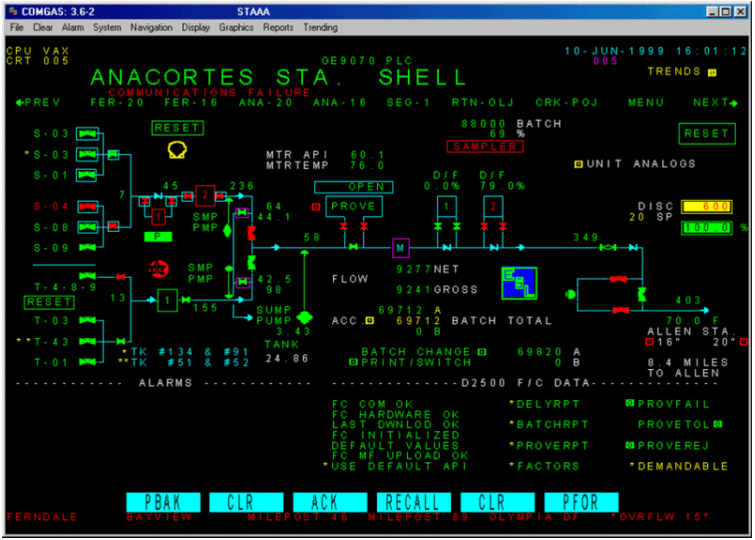
« Internet des Objets » (Internet of Things)

- Puces RFID partout
- Réseau électrique connecté, appareils ménagers connectés
- Robotique

Internet des Objets



Systèmes de contrôle



Moteurs du changement

- Alors : *big data* accessible partout, même en déplacement
- Outils d'analyse
  - Analyse/repérage comportemental
  - IA
- Chiffrement, anonymat

Moteurs du changement

Nouveaux endroits pour le *big data* :

- Le corps humain
  - Donnés de séquencage génétique
  - Biometrique (Apple Watch, Garmin, etc.)
  - Implants
  - Population vieillissante

Lecteurs de plaques

Voitures autonomes

Moteurs de changement

Moteurs économiques :

- Le coût de garder les données est maintenant plus bas que le coût de les supprimer
- Le coût de collectionner tout ce qu'on peut est maintenant plus bas que le coût d'être sélectif

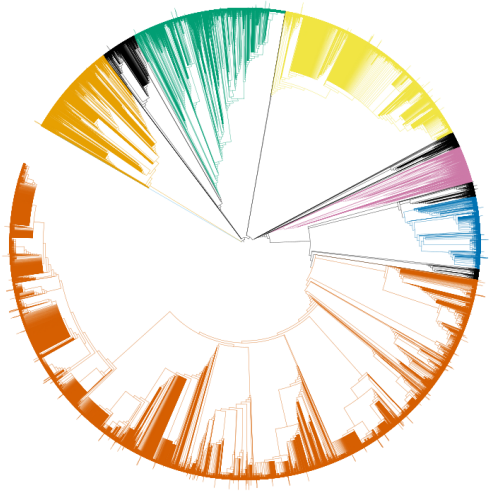
Changement dans le concept de « données »

- De questions de « création » aux questions d'« accès » et d'« attribution »
- La mort de *l'obscurité pratique* et de *l'anonymat automatique*
- Surveillance en continu et normalisée
- Données qui savent plus que nous

Le *big data* dans la science

Deux caractéristiques sont ici à retenir : la *taille* de ce répertoire, que l'on devine gigantesque ; le degré de *standardisation,* tel que des publications qui incluraient une espèce donnée…donnent lieu à une même liste de références, incluant photos, ensemble des localisations, date de chaque photo, caractéristiques de l'espèce, mention dans un tableau comprenant toutes les occurrences de cette espèce. (Huneman, p. 49)

Le *big data* dans la science



Les « voyages des données »

> *data journeys :* the material, social, and institutional circumstances by which data are packaged and transported across research situations, so as to function as evidence for a variety of knowledge claims. (Leonelli 2016, 5)

Les « voyages des données »

> Yet for successful reuse to take place, the journey that data ultimately undertake should be determined as much by their users as it is by their curators. Curators cannot possibly predict all the ways in which data might be used. This would involve familiarity with countless research programs… Therefore, the best way to explore and maximize the value of data as evidence is to enable as many researchers as possible to use data in their own way and within their own research context. (Leonelli 2016, 26)

Le *big data* dans la science

In October 2018, the CERN Data Centre passed the milestone of 300 petabytes of data permanently archived in its tape libraries. At the end of 2018, 330 PB of data were permanently archived on tapes in the CERN Data Centre.

New record in November 2018 for data taking over a single month : 15.8 petabytes of data (from all sources) were written on tape that month.

Within one year, when the LHC is running, more than one exabyte (the equivalent to 1000 petabytes) of data is being accessed (read or written). (CERN)

Le *big data* dans la science

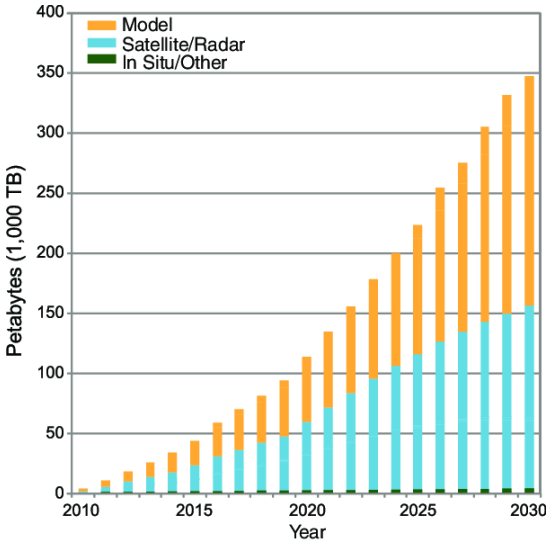
> Within one year, when the LHC is running, more than one exabyte (the equivalent to 1000 petabytes) of data is being accessed (read or written).

gigabyte → terabyte (1000 GiB) → petabyte (1000 TiB) → exabyte (1000 PiB)

Le *big data* dans la science

> Estimates predict that genomics research will generate between 2 and 40 exabytes of data within the next decade. (NHGRI)

Le *big data* dans la science



Le *big data* dans la science

Le *Big Data* ne diffère pas des simples données simplement en quantité ; c'est l'espace dans lequel ces données se déploient qui en vient à différer. Là où la donnée classique requiert un ou deux axes (taille, poids, dans mon exemple), elle peut en exiger des centaines dans le cas du *Big Data,* pour lequel l'espace des nuages de points où l'information pertinente doit être repérée devient un hyperespace à « n » dimensions, ce « n » pouvant indéfiniment croître. (Huneman, p. 54)

Faire de la science avec le *big data*

Très généralement, une distinction utile pour comprendre comment « fonctionne » une explication scientifique passe entre ce qu'on appelle en anglais *pattern,* patron ou gabarit en français, et les processus. Le *pattern* est une forme comme notre distribution d'arbres en forêt tropicale, mais ce peut être aussi bien n'importe quelle courbe mathématique selon laquelle s'ordonnent les données – et le processus, lui, génère ce *pattern,* donc son identification explique pourquoi il est là. (Huneman, p. 56)

Faire de la science avec le *big data*

> Ainsi, nous connaissons des relations causales dans la nature et la société, et elles expliquent les faits ; nous y avons accès en établissant des *patterns* de corrélation, mais à eux seuls ceux-ci ne garantissent pas un lien de causalité. Ces corrélations nous permettent souvent de prédire, mais la prédiction n'est pas toujours fiable. (Huneman, p. 58)

Faire de la science avec le *big data*

En écologie, ces dernières années un débat méthodologique a opposé certains chercheurs soutenant l'idée que la masse de données que nous récoltons, combinée au pouvoir nouveau du traitement statistique aidé par des algorithmes de type *machine learning,* permettrait de prédire les changements de la biodiversité ou de la composition des écosystèmes – par exemple, en regard du changement climatique – sans construire une modélisation explicitement causale des processus en jeu – autrement dit, sans *comprendre* ce qui se passe dans le système. (Huneman, p. 62)

Six questions pour le *big data* en science

1. Automatiser la recherche change la définition de la connaissance

> Il s'agit d'un monde où les quantités massives de données et les mathématiques appliquées remplacent tout autre outil susceptible d'être utilisé. Toutes les théories sur le comportement humain, de la linguistique à la sociologie, ont disparu. Autant pour toute théorie du comportement humain, de la linguistique à la sociologie. Oubliez la taxonomie, l'ontologie et la psychologie. Qui sait pourquoi les gens font ce qu'ils font ? [...] Avec suffisamment de données, les chiffres parlent d'eux-mêmes. (Anderson 2008)

Six questions pour le *big data* en science

1. Automatiser la recherche change la définition de la connaissance

Mais toute utilisation du *big data* apporte un tas de présupposés théoriques. Quelles sont leurs limitations ?

- Biaisé vers le présent (parfois mal archivé)
- Biaisé vers le texte (et contre image, vidéo, etc.)
- Biaisé vers le groupe (et contre l'indivdu)

Chaque méthode de recherche est limité ! Il faut **mieux comprendre comment ces aspects changent notre recherche.**

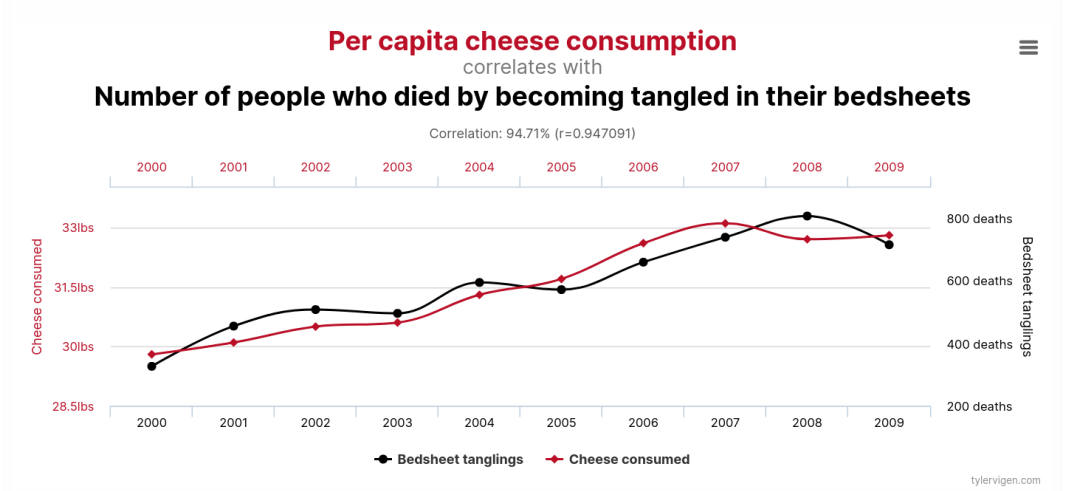
Six questions pour le *big data* en science

2. Déclarations de l'objectivité ou de l'exactitude sont souvent incorrectes

Il reste **toujours des jugements subjectifs** dans l'interprétation du *big data,* même si les données sont de qualité parfaite.

- Qu'est-ce qui compte comme une donnée, qu'est-ce qui sera compté ou mesuré ?
- Quelles variables sont du « bruit » qui doit être « ignoré » ?
- Comment gérer les erreurs dans nos données, ou compenser pour des données manquantes ? Qu'est-ce qu'une « aberration statistique » ?

Six questions pour le *big data* en science



Six questions pour le *big data* en science

3. Plus de données n'équivaut toujours aux meilleures données

Ajouter **plus de données** ne sert à rien si les données **ne sont pas une échantillon représentative de la population.**

Un bon exemple : analyser plus de Tweets de veut pas dire que l'image de « la politique belge » sur Twitter représente toute la population belge ! D'autant plus quand Twitter ne veut pas vous dire comment ils sélectionnent les tweets qu'ils fournissent aux chercheurs.

Six questions pour le *big data* en science

3. Plus de données n'équivaut toujours aux meilleures données

Si **plus de données** veut dire **intégrer des sources multiples,** ça entraîne d'autres problèmes.

Par exemple : pour arriver à une bonne perspective d'une population biologique, comment intégrer données sur la distribution dans l'espace et le temps, la génétique, les observations « classiques » du comportement, etc...

Six questions pour le *big data* en science

4. Hors contexte, le *big data* perd son sens

Les abstractions mathématiques ou statistiques qu'on prend du *big data* ne sont pas toujours équivalentes aux autres objets d'étude.

Exemple : Si l'on tire le « réseau » d'une personne sur Facebook ou Twitter, c'est pas clair du tout comment c'est lié aux autres réseaux qu'on a étudié dans le passé – des relations familiales, ou au travail, ou le réseau social « en présentiel ».

Six questions pour le *big data* en science

5. Être accessible ne veut pas dire être *éthique*

Le fait de pouvoir utiliser des données ne veut pas dire que les personnes qui les ont créées consentent à leur utilisation pour tel ou tel projet de recherche.

Comment assurer que les projets *big data* respectent l'éthique ? Les scientifiques peuvent savoir des choses à propos de nous que nous ne savons même pas.

Six questions pour le *big data* en science

6 Accès limité au *big data* crée nouvelles fractures numériques

Maintenant, l'accès à ces outils et, surtout, ces grandes bases de données, coût **très** cher. Est-ce que les projets de recherche là-dessus ne seront pas disponibles qu'aux chercheurs et chercheuses bien financé·e·s, des grandes universités de l'Europe et de l'Amérique du Nord ?

Six questions pour le *big data* en science

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Ouvrir la science !

LA SCIENCE OUVERTE | NOS ACTIONS | RESSOURCES | ACTUALITES | QUI SOMMES NOUS ?

Agir pour une recherche scientifique ouverte et partagée

Le Comité pour la science ouverte assure la mise en œuvre de la politique nationale de science ouverte.

→ **Découvrir le Plan national pour la science ouverte**

→ **Voir les projets du Comité**

Qu'est-ce que la science ouverte ?

Découvrez des guides, des recommandations et un lexique afin de vous initier à la science ouverte.

La question principale

À quel type de connaissance correspond un savoir prédictif sans compréhension de « ce qui se passe » ? Sommes-nous revenus au fond à ce que Platon voyait dans l'astronomie, une manière de « sauver les phénomènes », autrement dit de décrire les orbites et prédire les positions des corps célestes sans aucune connaissance sous-jacente des processus et des mouvements ? Ou bien s'agit-il de quelque chose d'épistémologiquement nouveau ? (Huneman, p. 67)