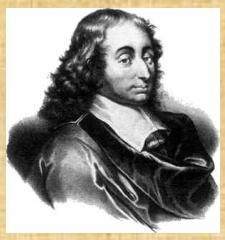
AFCSET

*Journées d'Andé 2017*


# *La représentation face à l'explosion des données*

## L'épistémologie de la représentation à l'ère du Big Data



***Jean-Paul Bois-Margnac***
**Vice-président de l'AFSCET**

***Notre propos :***

*examiner, à la lumière d'une réflexion*

*de nature épistémologique,*

*la pertinence des données, des faits et des profils*

*engloutis par le Big Data dans ses « Océans de données » …*

**Mais avant d'aborder la partie « épistémologie »**
**présentons quelques repères quantitatifs**
**sur le Big Data …**

## Le Big Data, des chiffres hallucinants …

… on en prend conscience avec l'inflation exponentielle des préfixes …

… où sont passés nos Mégas, Gigas, Terras ? …

Les volumes des données brassés par le Big Data s'expriment désormais en « Zetta octets » après avoir rapidement sauté les « Peta et les Hexa » !

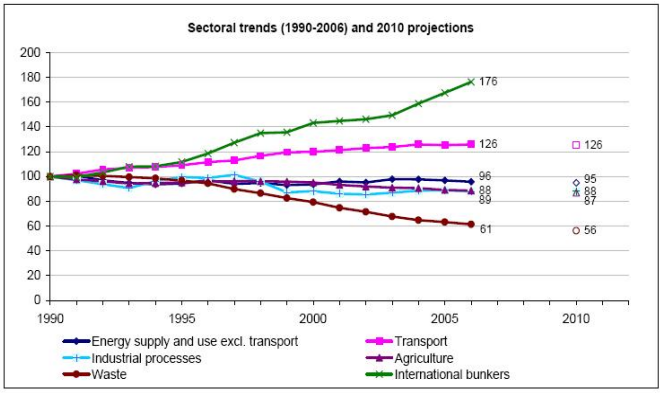
Pour mémoire : un Zetta octets = $10^{21}$ octets !

## L'unité de compte du Big Data est donc le Zetta octets !

Téra = $10^{12}$

Péta = $10^{15}$

Exa = $10^{18}$




Zetta = $10^{21}$ octets

*Notons que la production/consommation d'énergie ne connaît pas la même inflation et ne se compte qu'en Téra Watts …*

## D ’autres indicateurs de volumétrie …

- 1,2 zettaoctets par an en 2010
  - 1,8 zettaoctets en 2011,
    - 2,8 zettaoctets en 2012
      - elles s'élèveront à 40 zetta octets en 2020 !

En janvier 2013, Twitter générait 7 téraoctets chaque jour et Facebook, 10 téraoctets.
En 2014, Facebook générait déjà 4.000 To de data par jour.

Les installations scientifiques produisent encore plus de données.
Le radiotélescope “Square Kilometre Array”, produira 50 téraoctets de données analysées par jour, à un rythme de 7.000 téraoctets de données brutes par seconde.

## D’autres chiffres éloquents en termes de flux :

1992 : 100 Go/jour

1997 : 100 Go/heure

2002 : 100 Go/seconde

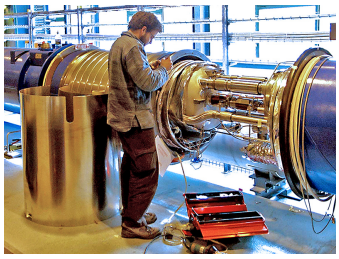
2013 : 28.875 Go/seconde

2018 =~50.000 Go/seconde …

-> Concept de « Vélocité » : la capacité de traiter en temps-réel un flux intense de données …

## Quelles applications ont engendré cette inflation de données ?

Les réseaux sociaux : Twitter, Facebook …



Les applications scientifiques …

La diffusion d'images : Flickr, YouTube, …

Les messageries …

Les jeux vidéos …

« Pass » des transports publics …

Mais aussi les nombreux automates …

Péages sans contacts …

Compteurs EDF « Intelligents » …

… qui contribuent à alimenter le Big Data …

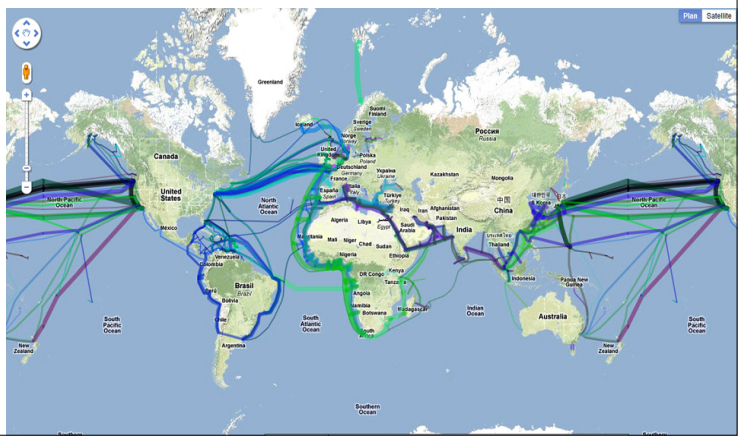
**Pour stocker et traiter ces océans de données,**
**de gigantesques Data Centers sont implantés un peu partout …**


**Sur ce le toit de ce centre, les panneaux solaires fournissent 70 megawatts en crête !**

**Aujourd'hui, les capacités des Data Centers semblent infinies …**

**Leur *interconnexion* et leur *interopérabilité* ont créé**

**ce que l'on appelle désormais le Cloud …**



## *Où l'on évoque déjà un cinquième pouvoir …*

**Le continent du Big Data est essentiellement constitué des GAFA (Google, Apple Facebook, Amazon) …**

**… et les américains contrôlent déjà 72% des cinquante plus grands sites mondiaux qui l'alimentent …**

**Gilles Babinet, dans un rapport de 2015 pour l'Institut Montaigne, rappelait qu'Amazon investit 1,4 milliards de $ par an dans son Cloud …**

## Alors, quelle métaphore est la mieux à même de représenter le phénomène ?

**… une pieuvre ?**

**un « nuage »**

**ou …**

**Cette interrogation sémantique n'est pas anodine …**

**… un nuage est évanescent,**

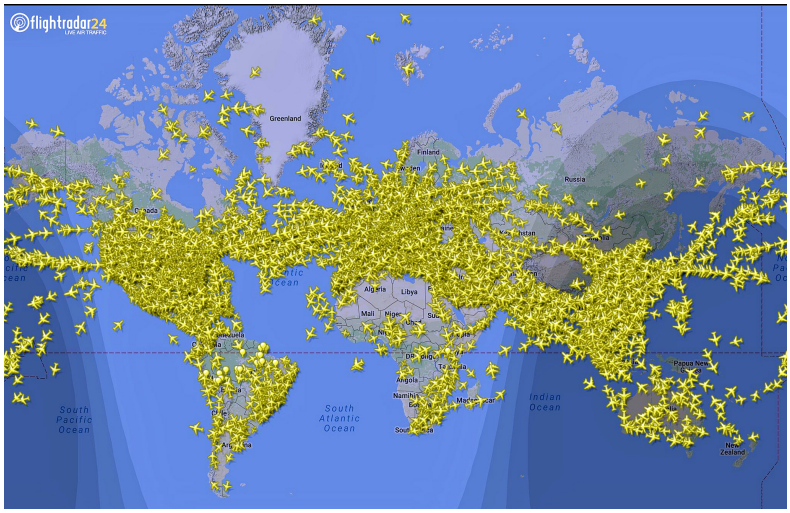
**on ne sait d'où il vient, on ne peut le saisir …**

**… dans le « nuage », les responsabilités**

**sont diluées …**

**… une image plus concrète, plus biologique**
**semble mieux appropriée …**

**Mais le Big Data, c'est aussi des outils**
**à la mesure du « Village Global » …**




**Sur ce planisphère, l'application FlightRadar suit**
**plus de 6.000 vols en temps-réel !**

**Le versant « privacy » du Big Data est plus inquiétant …**

**Au-delà des données scientifiques ou techniques,**
**la capture de données personnelles fait peser**
**de sérieuses menaces sur la vie privée**
**des citoyens du monde entier …**

***Le syndrome d'un Big Brother puissance dix !***

**Si le projet SAFARI d'interconnexion des fichiers**
**du Min. de l'Intérieur avait suscité en 1974**
**une légitime levée de boucliers et conduit à l'adoption**
**de la Loi Informatique et Liberté en 1978, …**

***… le Big Data, de par sa distribution planétaire,***
***semble échapper à toute régulation …***

***… d'où l'intérêt de se poser la question de la pertinence***
***des données/informations stockées et surtout***
***de leur traitement algorithmique …***

*L'inquiétude des citoyens vient en fait plus du*

*traitement algorithmique des données*

*que de la captation des données elles-mêmes …*

*… inquiétude renforcée par le fait que la*

*majorité des critères utilisés par les algorithmes*

*ne sont pas connus (e.g. credit scoring, « amis » sur FB …)*

**Une illustration de cet effet pervers :**

*Un chercheur en sociologie prépare*

*une thèse sur l'homosexualité masculine.*

*Il utilise Facebook comme source d'information …*

*Le réseau social lui proposera de plus en plus « d'amis »*

*homosexuels car ses algorithmes l'auront*

*très probablement « profilé »*

*comme homosexuel lui-même …*

**Devant le côté intrusif de ces algorithmes,**

**il est donc légitime de se poser quelques questions …**

*Mais pourquoi proposer un détour par*

***l'épistémologie** pour examiner cette menace*

*d'une surveillance purement **algorithmique** ?*

*L'épistémologie, la **théorie philosophique***

***de la connaissance,** pose peut-être*

*la bonne question …*

**Comment connais-tu ce que tu crois connaître ?**

**Donc, appliqué au Big Data,**

**que ses outils sont-ils en mesure de connaître ?**

*En fait, cette pertinente interrogation*
*recèle deux questions :*

**Comment connais-tu ?**

*** par quelles voies**

*** à l'aide de quelles représentations ?**

**Que crois-tu connaître ?**

*** Es-tu certain que l'objet de ta connaissance**
**n'échappe pas à tes capacités de représentations ?**

**A propos de la première question,**
**usons de cette image :**

**A quoi sert de posséder une bibliothèque renfermant**

**tous les faits et toutes les connaissances**

**si l'on ne sait ni où se trouvent les ouvrages ni,**

***a fortiori*, comment les lire …**

**Pour le Big Data, la question « par quelles voies connais-tu » se rapporte donc aux outils successivement mis au point pour traiter de grandes masses de données :**

***Des plus rustiques aux plus sophistiqués :***

- **SGBD** (Relationnelles, OO) **60'-70'**
- **Data Mining 80'-90'**
- **Deep Learning > 2.000**

***Une efficacité croissante dans le traitement des données …***

**Bases et Banques de données**

→ **Croisement de fichiers …**

**Data Mining**

→ **Corrélations entre faits et données…**

**Deep Learning**

→ **Les précédents + la fusion avec l'IA…**

**Le Deep Learning ou Apprentissage Profond**
tire profit des avancées de l'**Intelligence Artificielle**
pour conférer aux machines des capacités d'apprentissage
par l'examen d'un nombre colossal de cas et,

- soit d'assurer **une modélisation à un haut niveau d'abstraction,**
- soit de **reconnaître une scène** dans un environnement inconnu…

**Exemples fameux :**

- **le programme AlphaGo de Google a battu les meilleurs joueurs de Go …**
- **la voiture autonome …**

**Pour Yann Le Cun** (FaceBook, Collège de France),
**l'un des inventeurs du Deep Learning.**

***« Avant, il fallait…expliquer à l'outil comment transformer une image afin de la classifier.***

***Avec le deep learning, la machine apprend à le faire elle-même.***

***Et elle le fait beaucoup mieux que les ingénieurs,***

***c'est presque humiliant !»***

**Comment connais-tu ?**

*** par quelles voies**

*** à l'aide de quelles représentations ?**

- **Octets (ASCII) ->** mots isolés, phrases, contexte …
  - **Pixels** –> Voxels -> Images fixes et animées

**Là encore, la complexification des représentations accroît drastiquement la puissance de révélation des algorithmes …**

**Revenons maintenant à notre interrogation :**

**Comment connais-tu ce que tu crois connaître ?**

**Ou, formulé plus précisément, comment le Big Data connaît-il ce qu'il croit connaître ?**

## Aller sur le Net, c'est comme marcher sur la neige …

## …quoi qu'on fasse, on y laisse des traces …

## On estime à plus de 3 milliards le nombre de personnes ayant accès à Internet sur la planète …

## Autant de sources de données alimentant le Big Data …

« Pass » des
transports publics …

Et, comme nous l'avons déjà souligné, de nombreux automates autoes …

Péages sans contacts …

… contribuent aussi à nous suivre à la trace …

Compteurs EDF « Intelligents » …

Le croisement de toutes les données englouties par le Big Data dans ses gigantesques « entrepôts de données » nous *profile* :

-> préférences sexuelles

-> opinions politiques

-> addictions de tous genres

-> habitudes consuméristes

-> la liste est sans fin …

D’où un certain nombre

de craintes raisonnées

et raisonnables …

## Et alors ?

Posons-nous cette question :

que savent réellement les « profileurs »

et quel pouvoir cela leur confère-t-il

et à quelles fins ?

## Chaque être humain cumule plusieurs identités …

- Citoyenne
- Financière/consumériste
- Professionnelle
- Familiale
- Sexuelle …

**Pour l'instant, aucun algorithme ne semble en mesure de les fondre en une seule qui les résumerait toutes …**

**Dans un Etat de Droit et, compte-tenu des technologies disponibles, ces profilages contribuent surtout à optimiser des ressources et à prendre des décisions …**

- Physique/astrophysique
- Finance
- Logistique
- Energie
- Marketing
- Médecine …

**L'aspect le plus visible du Big Data pour le citoyen est cette sensation désagréable en allant, par exemple, sur Amazon, que le site sait tout de ses goûts …**

## Vers l'émergence d'un citoyen 2.0 ?

-> plus prudent dans la protection de sa « privacy » ?

e.g. « Navigation privée » systématique…

-> entraîné à ne pas se laisser déstabiliser

e.g. être conscient que l'on peut nous opposer des faits très anciens

-> exigeant un véritable « droit à l'oubli » !

e.g. procédures simplifiées pour effacer nos traces …

## En 1645 la « Pascaline » …

… libérait les comptables des fastidieuses opérations arithmétiques …

**Quatre siècles plus tard,**
**sa descendance technologique …**

**… mettrait en œuvre**
**un asservissement digital**
**encore plus intrusif que celui**
**imaginé par George Orwell ?**


Personnage imaginé
par le street artist
Shepard Fairey

**Peut-on quand même terminer**
**sur une note plus optimiste ?**

**Pas vraiment !**

## Le véritable Big Brother pourrait venir de la coopération entre le Deep Learning et la magnéto-encéphalographie …

Cette technologie d'imagerie cérébrale fonctionnelle
permet de mesurer l'activité cérébrale
en temps réel et d'établir quand, où, et comment
le cerveau nous confère la sensation, la perception,
la mémoire, le langage, la pensée,
les processus de décision, et le contrôle des actions …

**En un mot, et ce n'est pas de la science-fiction, de lire nos pensées.**

## A l'avenir, et plus que jamais,
## il faudra donc des citoyens vigilants
## et des régulations efficaces !

# *Merci* . . .

www.afscet.asso.fr jeanpaul.bois@gmail.com