

# Big Data

## Plan

1. Introduction
2. Analyse des grands réseaux d'interactions

# Ressources

→ Pour me joindre

`mailto: olivier.michel@u-pec.fr`

→ Sur mon site web

`http://www.lacl.fr/~michel/doku.php?id=teaching:bigdata:start`

→ Avec le login/passwd donné en cours

- Support de cours
- Articles
- Autre documents

→ Pourquoi cette police de caractère : `http://bit.ly/2mAjtKH`

# Contrôle des connaissance

- Un contrôle continu
- Un projet



Pour ceux qui le souhaitent : lecture/exposé d'articles (en +)

# Références bibliographiques et sources

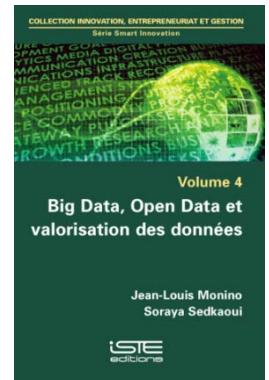
- *Les Big Data à découvert* - Sous la direction de M. Bouzeghoub & R. Mosséri - CNRS Éditions - 2017



- Le Big Data - Pierre Delort - PUF - 2015



- Big Data, Open Data et valorisation des données - Jean-Louis Monino & Soraya Sedkaoui - ISTE Éditions - 2016



- ...et de très nombreux articles scientifiques [disponibles sur demande]

# Objectifs du cours

## Première partie

1. Comprendre les enjeux associés au Big Data (cours)
2. Connaitre les éléments historiques, théoriques et pratiques du Big Data (cours)
3. Comprendre et maitriser une méthode théorique d'analyse des données (cours)
4. Extraire, filtrer et analyser un grand réseau d'interaction (TP)
5. Découvrir un outil de visualisation interactive des grands graphes (TP)

## Seconde partie

1. Comprendre et maitriser la technologie MapReduce et l'environnement Hadoop (cours + TP)
2. Mettre en œuvre les TP dans un environnement de virtualisation VMware vSphere

# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# De l'industrialisation d'Internet...

## Rise of the Machines

WHAT HAPPENED WHEN  
**1B PEOPLE**  
BECAME CONNECTED?

Entertainment is Digitized  
Social Marketing Emerged  
Communications Mobilized  
IT Architecture Virtualized  
Retail & Ad Transformed



WHAT HAPPENS WHEN  
**50B Machines**  
BECOME CONNECTED?

Operational Tech (OT) is virtualized  
Analytics become predictive  
Machines are self healing & automated  
Monitoring and maintenance is mobilized  
Employees increase productivity



<https://www.youtube.com/watch?v=qXYZDd2heK8>

# ...au déluge des données



<https://www.youtube.com/watch?v=0Q3sRStUymS>

# Big Data

## Plan

### 1. Introduction

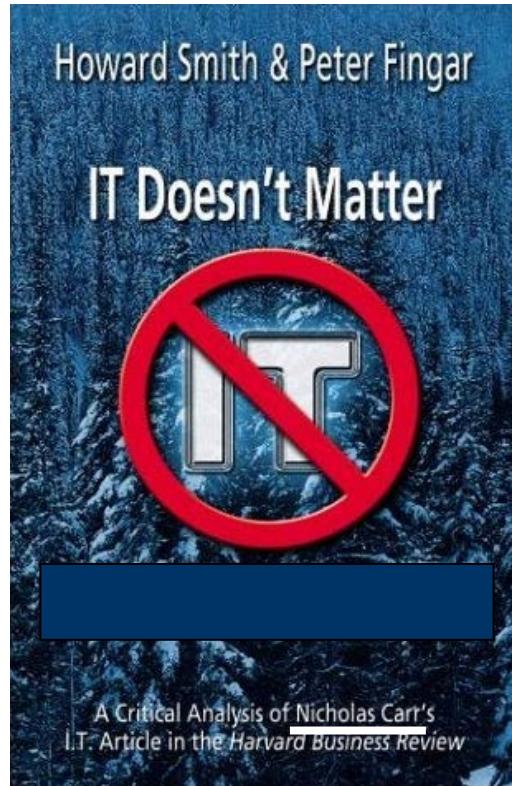
- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# À trop se focaliser sur les traitements...

→ N. G. Carr, "IT Doesn't Matter", *Harvard Business Review*, 2003.



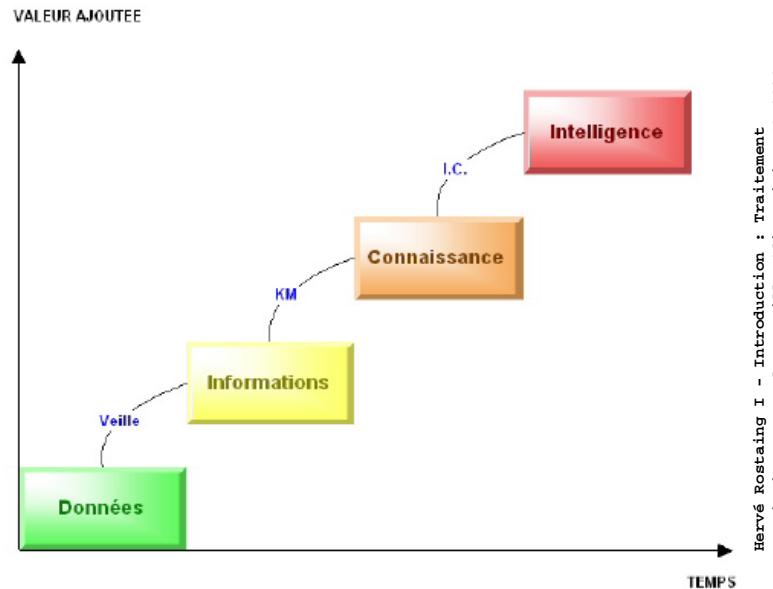
→ Tout le monde dispose d'ERP. Il faut investiguer de nouveaux champs fonctionnels !

# Le Big Data

- Cet obscur objet du désir... de quoi s'agit-il ?

# Le Big Data

- Cet obscur objet du désir... de quoi s'agit-il ?
- L'information (interprétation de symboles) ?



# Le Big Data

- Cet obscur objet du désir... de quoi s'agit-il ?
- ☒ L'~~information (interprétation de symboles)~~ ?
- ➔ La **donnée** (valeur d'une variable/capteur)

# Le Big Data

- Cet obscur objet du désir... de quoi s'agit-il ?
- ☒ L'~~information (interprétation de symboles)~~ ?
- ➔ La donnée (valeur d'une variable/capteur)
- ➔ L'analyse de données ? Le BI alors ?

# Le Big Data

- Cet obscur objet du désir... de quoi s'agit-il ?
- ☒ L'~~information (interprétation de symboles)~~ ?
- ➔ La donnée (valeur d'une variable/capteur)
- ➔ L'analyse de données ? Le BI alors ?
-  Non, car alors l'exploration des données s'appuie sur le *modèle structurant les données (Datawarehouse)*

# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- [Les deux catégories de raisonnement](#)
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# Deux types de raisonnements

## □ Déductif (logique top-down)

- Un système de règles d'inférences
- Des hypothèses
- On en conclut des faits

→ Modèle *a priori* (maths, physique...)

## □ Inductif (logique bottom-up)

- On dispose de données issues de l'observation
- On est capable de corrélérer ces données ensemble
- On en déduit que ces données (conclusions) découlent d'hypothèses avec une certaine probabilité

→ Modèle *a posteriori* (biologie, shs, sciences expérimentales, Big Data...)

Règles de la déduction naturelle

Hypothèse	Classique	Coupe
$A \in \Gamma$ $\frac{}{\Gamma \vdash A}$	$\frac{\Gamma \vdash \neg A}{\Gamma \vdash A}$	$\frac{\Gamma, A \vdash B \quad \Gamma \vdash A}{\Gamma \vdash B}$
Introduction( $I$ )	Elimination( $E$ )	Hypothèse( $H$ )
$\top$ $\frac{}{\Gamma \vdash \top}$	$\frac{\Gamma \vdash \perp}{\Gamma \vdash C}$	$\frac{\Gamma \vdash A}{\Gamma, \top \vdash A}$
$\perp$ $\frac{}{\Gamma, A \vdash \perp}$ $\frac{}{\Gamma \vdash \neg A}$	$\frac{\Gamma \vdash \neg A \quad \Gamma \vdash A}{\Gamma \vdash C}$	$\frac{}{\Gamma, \perp \vdash C}$ $\frac{\Gamma, \neg A \vdash A}{\Gamma, \neg A \vdash C}$
$\wedge$ $\frac{\Gamma \vdash A \quad \Gamma \vdash B}{\Gamma \vdash A \wedge B}$	$\frac{\Gamma \vdash A \wedge B}{\Gamma \vdash A}$ $\frac{\Gamma \vdash A \wedge B}{\Gamma \vdash B}$	$\frac{\Gamma, A, B \vdash C}{\Gamma, A \wedge B \vdash C}$
$\vee$ $\frac{\Gamma \vdash A \quad \Gamma \vdash B}{\Gamma \vdash A \vee B}$ $\frac{\Gamma \vdash B \quad \Gamma \vdash A \vee B}{\Gamma \vdash A \vee B}$	$\frac{\Gamma \vdash A \vee B \quad \Gamma, A \vdash C \quad \Gamma, B \vdash C}{\Gamma \vdash C}$	$\frac{\Gamma, A \vdash C \quad \Gamma, B \vdash C}{\Gamma, A \vee B \vdash C}$
$\Rightarrow$ $\frac{\Gamma, A \vdash B}{\Gamma \vdash A \Rightarrow B}$	$\frac{\Gamma \vdash A \Rightarrow B \quad \Gamma \vdash A}{\Gamma \vdash B}$	$\frac{\Gamma, B \vdash C \quad \Gamma, A \Rightarrow B \vdash A}{\Gamma, A \Rightarrow B \vdash C}$
$\forall$ $\frac{\Gamma \vdash P \quad x \notin V(\Gamma)}{\Gamma \vdash \forall x, P}$	$\frac{\Gamma \vdash \forall x, P}{\Gamma \vdash P[x \leftarrow t]}$	$\frac{\Gamma, (\forall x, P), P[x \leftarrow t] \vdash C}{\Gamma, (\forall x, P) \vdash C}$
$\exists$ $\frac{\Gamma \vdash P[x \leftarrow t]}{\Gamma \vdash \exists x, P}$	$\frac{\Gamma \vdash \exists x, P \quad \Gamma, P \vdash C \quad x \notin V(\Gamma, C)}{\Gamma \vdash C}$	$\frac{\Gamma, P \vdash C \quad x \notin V(\Gamma, C)}{\Gamma, (\exists x, P) \vdash C}$
$=$ $\Gamma \vdash t = t$	$\frac{\Gamma \vdash t = u \quad \Gamma \vdash P[x \leftarrow t]}{\Gamma \vdash P[x \leftarrow u]}$	$\frac{\Gamma, t = u \vdash P[x \leftarrow t]}{\Gamma, t = u \vdash P[x \leftarrow u]}$

# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- [L'exemple Google Flu](#)
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# L'exemple Google Flu - Évolutions de la grippe

- 250 000 à 500 000 décès par an  
(grippe espagnole de 1918-1920 : plusieurs dizaines de millions de morts)
- Virus ARN, composé de 12 gènes, sujet à mutations
- Deux causes aux mutations :
  - (1) glissement antigénique
  - (2) réassortiment génétique
- Le point (1) est pris en compte par la préparation d'un vaccin basé sur les souches virales de l'année (n-1)
- Le point (2) nécessite de déterminer *le virus en cours* pour permettre la mise en œuvre d'un vaccin

# L'exemple Google Flu - Gagner du temps

- Commence alors une course contre la *diffusion du virus* :
  - Perdue en 2009, souche H1N1, délai de 6 mois depuis le début de la pandémie
  - Vaccin disponible *après* le sommet de la seconde vague de la pandémie
- Comment gagner du temps :
  - ✓ Lors de la production du vaccin : délai réduit à 4 jours et 4 heures !
  - ➔ Lors de la détection de la souche virale en cours
- Accélérer la détection de la souche
  - Par un réseau de médecins qui maille le territoire
    - FR : réseau Sentinelles : 1 500 médecins - 2,2% des médecins généralistes
    - US : réseau Sentinel : 2 500 médecins, 9 zones - 12 m de consultations/an  
CDC analyse les données : **délai de 2 semaines** / à l'événement physique
  - Par des méthodes alternatives
    - Appels téléphoniques sur un numéro dédié
    - Consultation de sites web
    - Ventes de médicaments
    - **Recherches faites par les patients sur un moteur de recherche (Google)**

# L'exemple Google Flu - Un modèle inductif

## □ Google Flu

- Équipe mixte Google & CDC
- Basée sur les logs du moteur de recherche 2003-2007
- Publication dans Nature 2009

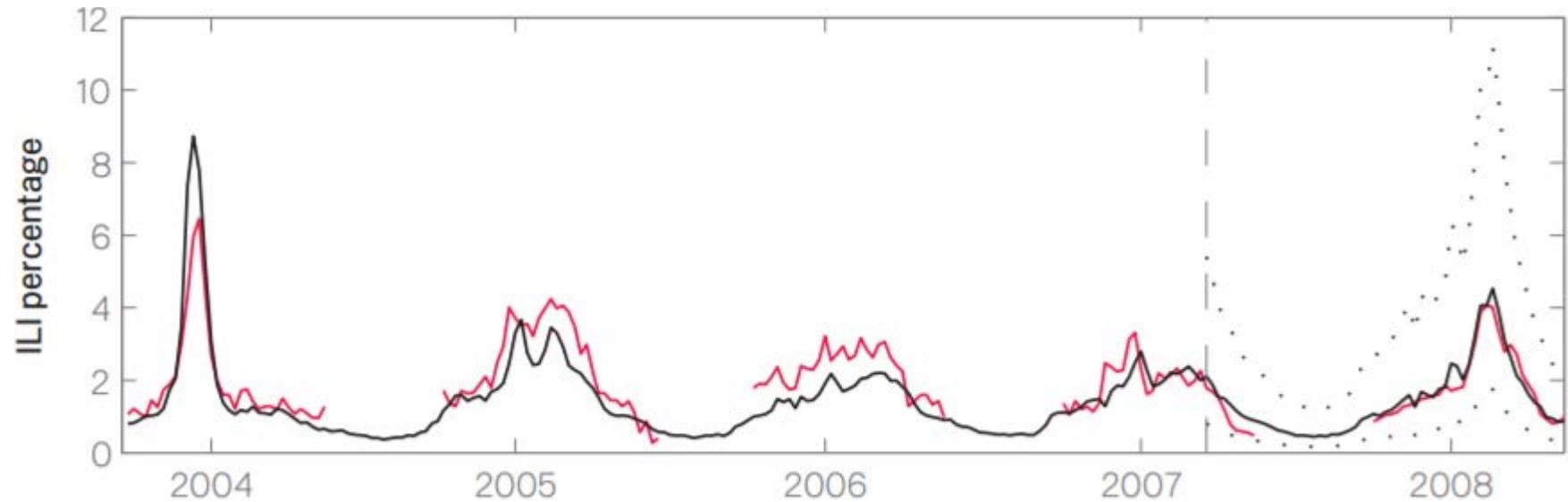
[Jeremy Ginsberg<sup>1</sup>, Matthew H. Mohebbi<sup>1</sup>, Rajan S. Patel<sup>1</sup>, Lynnette Brammer<sup>2</sup>, Mark S. Smolinski<sup>1</sup> & Larry Brilliant<sup>1</sup> - Detecting influenza epidemics using search engine query data - Nature 457, 1012-1014, 19/02/2009] (\*)

- Analyse de 50 millions de recherches, par état - 9 états
- Comparaison des termes de recherches ayant la meilleure corrélation avec les données du CDC
- Maximum obtenu pour 45 termes :

Search Query Topic	Top 45 Queries		Next 55 Queries	
	N	Weighted	N	Weighted
Influenza Complication	11	18.15	5	3.40
Cold/Flu Remedy	8	5.05	6	5.03
General Influenza Symptoms	5	2.60	1	0.07
Term for Influenza	4	3.74	6	0.30
Specific Influenza Symptom	4	2.54	6	3.74
Symptoms of an Influenza Complication	4	2.21	2	0.92
Antibiotic Medication	3	6.23	3	3.17
General Influenza Remedies	2	0.18	1	0.32
Symptoms of a Related Disease	2	1.66	2	0.77
Antiviral Medication	1	0.39	1	0.74
Related Disease	1	6.66	3	3.77
Unrelated to Influenza	0	0.00	19	28.37
	<b>45</b>	<b>49.40</b>	<b>55</b>	<b>50.60</b>

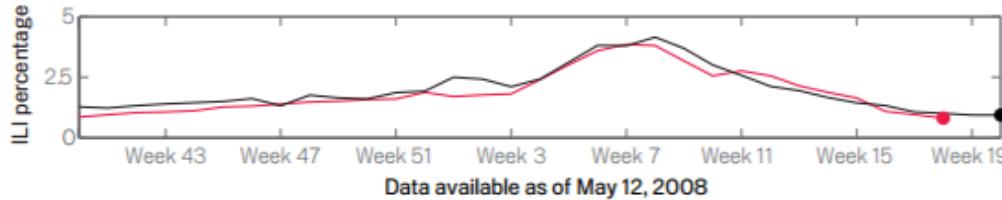
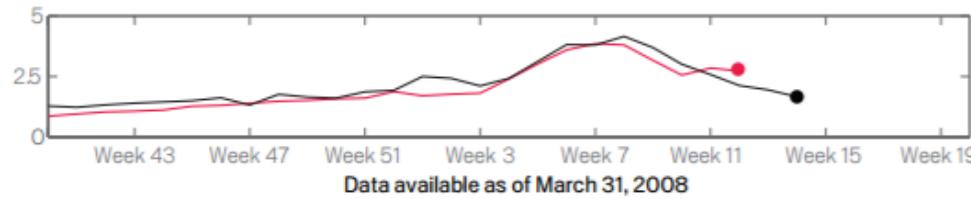
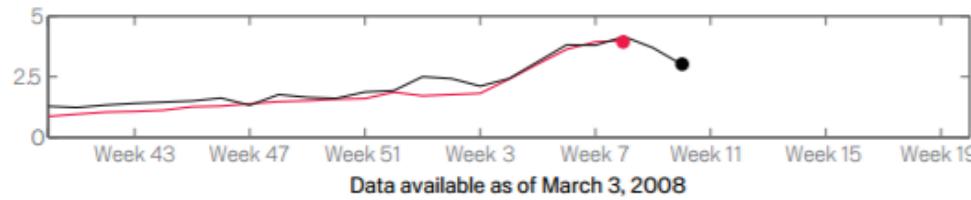
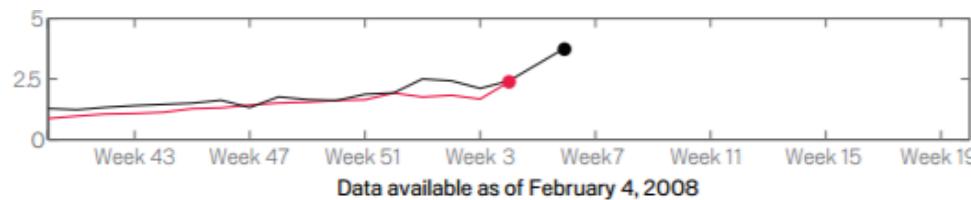
(\*) Toutes les données/graphiques qui suivent proviennent de l'article de Nature

# L'exemple Google Flu - Qualité du modèle



- Comparaison entre les estimations du modèle pour la région Mid-Atlantic (noir) et les données du CDC-ILI (rouge)
- Corrélation de .85 sur 128 points de cette région
- Corrélation de .96 sur 42 points

# L'exemple Google Flu - Prédiction

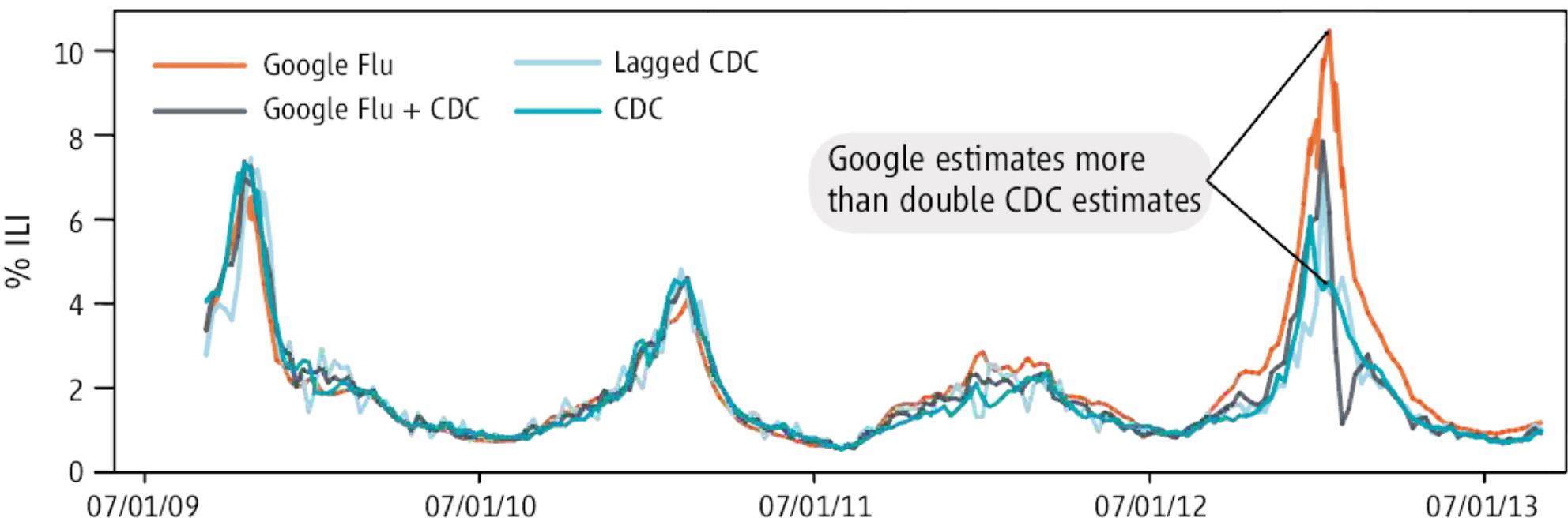


- Comparaison prédictions du modèle (noir) et données CDC (rouge)
- 2 semaines d'avance sur les données du CDC

# L'exemple Google Flu - Analyse

- On ne peut se passer des données réelles du CDC pour « caler » le modèle basé sur le moteur de recherche
- Limites du modèle
  - Si un changement de comportement a lieu, alors tout le modèle est faussé (habitudes de recherche, recherche faite par des tiers...)
  - Faible précision géographique (renforcée avec les terminaux mobiles)
  - Nécessite une importante population (pour lisser les variations)
  - Faux positifs (p. ex rappel d'un médicament antigrippal)

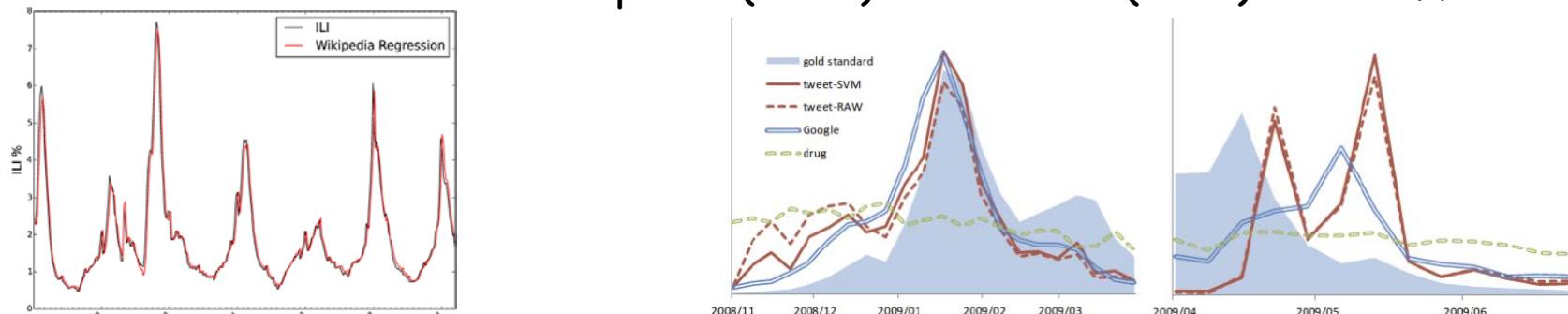
# L'exemple Google Flu - Analyse



- Limites atteintes lors de l'épisode fin 12/début 13
  - Souche plus virulente (H3N2)
  - Commence tôt (nov) et pic après noël
  - Le modèle prédit le double d'infections/médecins au 01/2013 à NY
  - Pas de commentaire de la part de Google

# L'exemple Google Flu - Critique

- Un modèle inductif nécessite des précautions
  - Apprécier l'erreur commise par le modèle dans sa représentation du monde (dépend de l'échantillon - taille, représentativité...)
  - Identifier les changements survenus depuis la création du modèle (complétion asynchrone, algo de recherche...)
  - Identifier les manipulations du modèle (entreprise pharmaceutique qui augmente ses requêtes pour faire remonter son produit)
  - Peu de transparence de Google Flu sur données/algos
  - Un travail similaire sur Wikipedia (2015) et Twitter (2011) a été effectué



# L'exemple Google Flu - Conclusion

- Début 2013, GF adopté par 29 pays
- Extension à la pathologie de la dengue
- Conservation des données depuis (au moins) 2003 chez Google
- Aggrégation de 50 millions de recherches / 9 états (2.5 M !)
- « *Big Data : créer en exploratoire et par induction sur des masses de données à faibles densité en information des modèles à capacité prédictives* »

[Def de P. Delort - Big Data - Que sais-je]

# Big Data

## Plan

### 1. Introduction

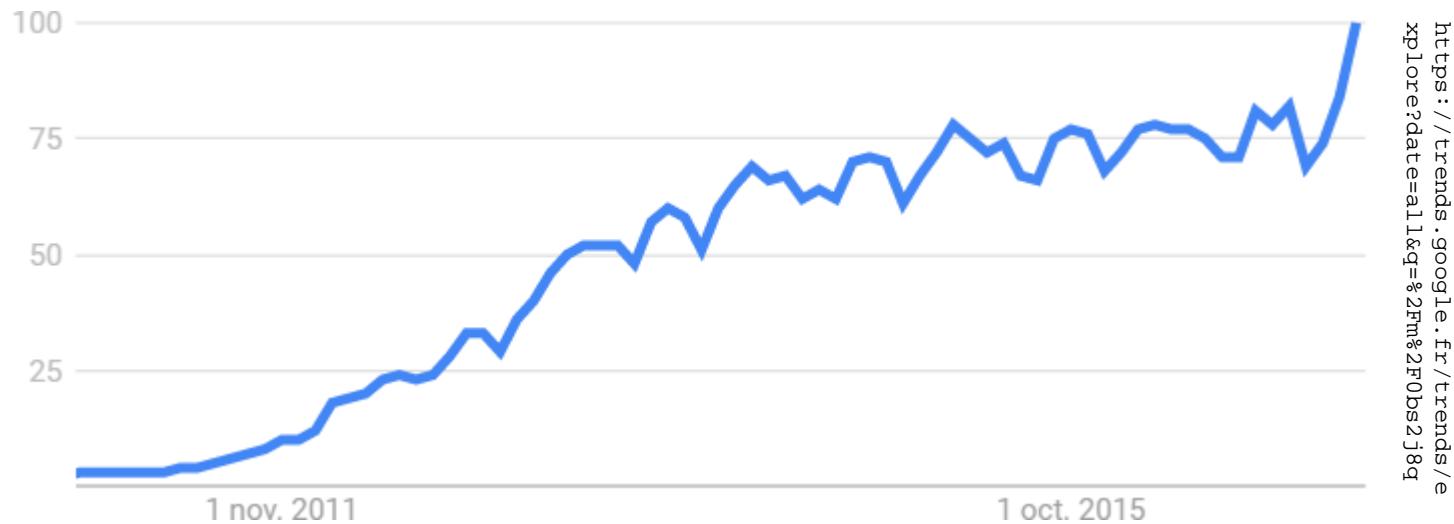
- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- [Histoire](#)
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# Google trends sur 'big data'

- Le Big Data, un sujet « important » ?



# Histoire du Big Data

- 1997, The problem of Big Data, NASA researchers, Michael Cox et and David Ellsworth's paper 
- 1998, Google was founded 
- 1999, Apache Software Foundation (ASF) was established 
- 2000, Doug Cutting launched his indexing search project: Lucene **  
2000, L Page and S. Brin wrote paper "the Anatomy of a Large-Scale Hypertextual Web search engine"
- 2001, The 3Vs, Doug Laney's paper "3D data management: controlling data Volume, Velocity & Variety" **
- 2002, Doug Cutting and Mike Cafarella started Nutch, a subproject of Lucene for crawling websites **
- 2003, Sanjay Ghemawat et al. published "The Google File System"(GFS)**  
2003, Cutting and Cafarella adopted GFS idea and create Nutch Distribute File System (NDFS) later, it became HDFS  
 
- 2004, Yonik Seeley created Solr for Text-centric, read-dominant, document-oriented & flexible schema search engine **  
2004, Google Began to develop Big Table 
- 2004, Jeffrey Dean and Sanjay Ghemawat published "Simplified Data Processing on Large Cluster" or MapReduce**  
2005 Nutch established Nutch MapReduce  
2005, Damien Katz created Apache CouchDB (Cluster Of Unreliable Commodity Hardware), former Lotus Notes 
- 2006, Cutting and Cafarella started Hadoop or a subproject of Nutch **  
2006, Yahoo Research developed Apache Pig run on Hadoop 
- 2007, 10gen, a start-up company worked on Platform as a Service (PaaS). Later, it became MongoDB **  
2007, Taste project  
2008, Apache Hive (extend SQL), HBase (Manage data) and Cassandra(Schema free) to support Hadoop   
2008, Mahout, a subproject of Lucene integrated Taste 
- 2008 Hadoop became top level ASF project**  
**2008 TUB and HPI Initiated Stratosphere Project and later become Apache Flink**   
2009, Hadoop combines of HDFS and MapReduce. Sorting one TB 62 secs over 1,460 nodes
- 2010, Google licenced to ASF Hadoop **  
2010, Apache Spark , a cluster computing platform extends from MapReduce for in-memory primitives 
- 2011, Apache Storm was launched for a distributed computation framework for data stream **  
2012, Apache Drill for Schema-Free SQL Query Engine for Hadoop, NoSQL and cloud Storage 
- 2012, Phase 3 of Hadoop – Emergence of "Yet Another Resource Negotiator"(YARN) or Hadoop 2**  
2013 Mesos became a top level Apache project 
- 2014, Spark has > 465 contributors in 2014, the most active ASF project   **  
2015, Enter Zeta Byte Era 

# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- **Les multiples Vs**
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# Les multiples V du Big Data (1)

→ Réminiscence des 3C (Customer/Competition/Challenge) de K. Ohmae

□ 3Vs du Gartner (2001) - D. Laney du Meta Group

1. Volume, which means Incoming data stream and Cumulative volume of data  
[For Intel median volume > 300 To/week]
2. Velocity, which represents the pace data used to support interaction and generated by interactions
3. Variety, which signifies the variety of incompatible and inconsistent data formats and data structures

□ 4Vs d'IBM

4. Veracity implies uncertainty of data

□ 6 Vs de Microsoft

5. Variability refers to the complexity of data set. In comparison with "Variety" (or different data format), it means the number of variables in data sets.
6. Visibility emphasises that you need have a full picture of data in order to make informative decision.

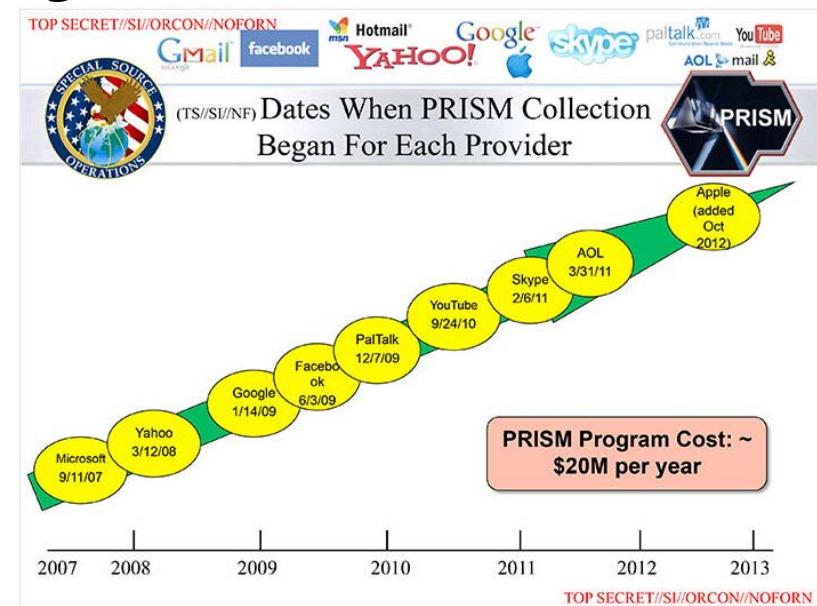
# Les multiples V<sup>illomies (?)</sup> du Big Data (1bis)

- 10 Juin 2013 : révélations d'E. Snowden sur PRISM/NSA
- Les 3 V, ne seraient-ils pas du *Big Business* ?

- Measure
- Manipulate
- Monetize

et du *Big Brother* ?

- Dataification
- Dataism
- Dataveillance



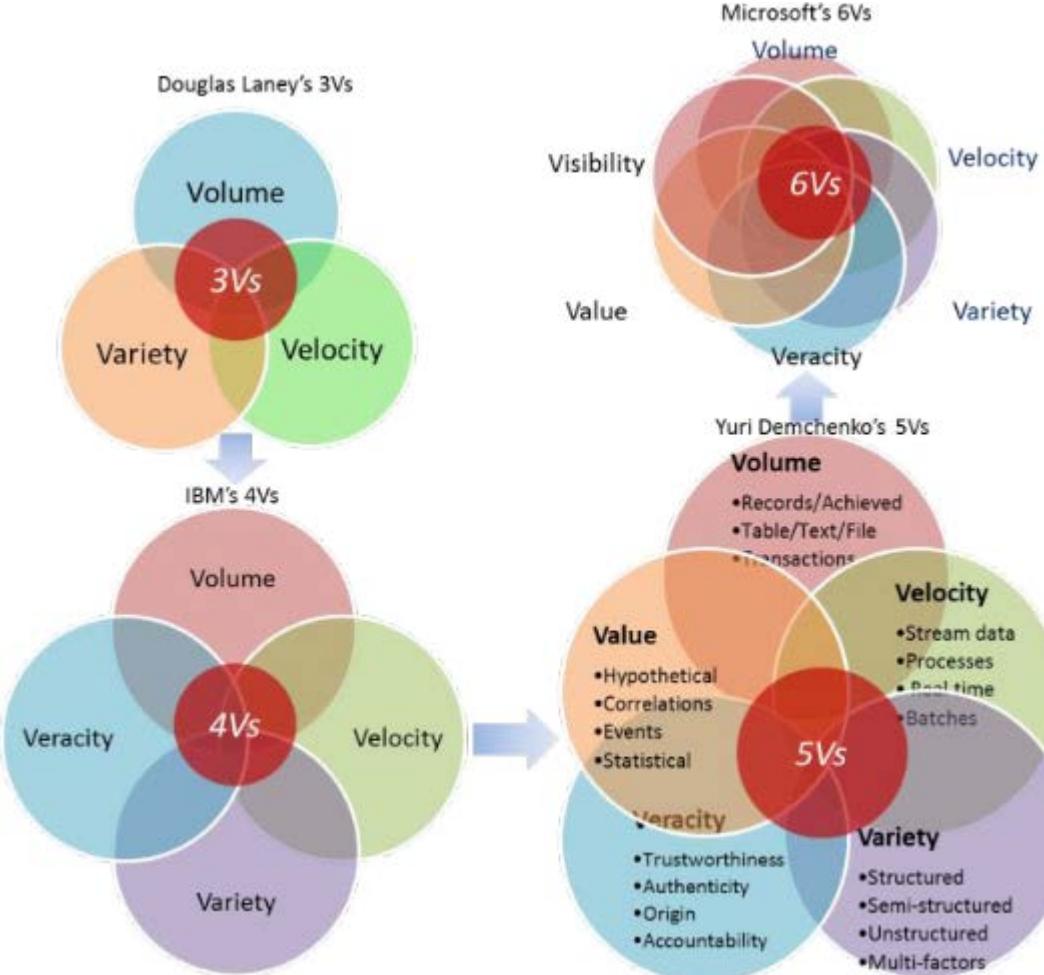
Max Kelly, responsable sécurité FB embauché à la NSA



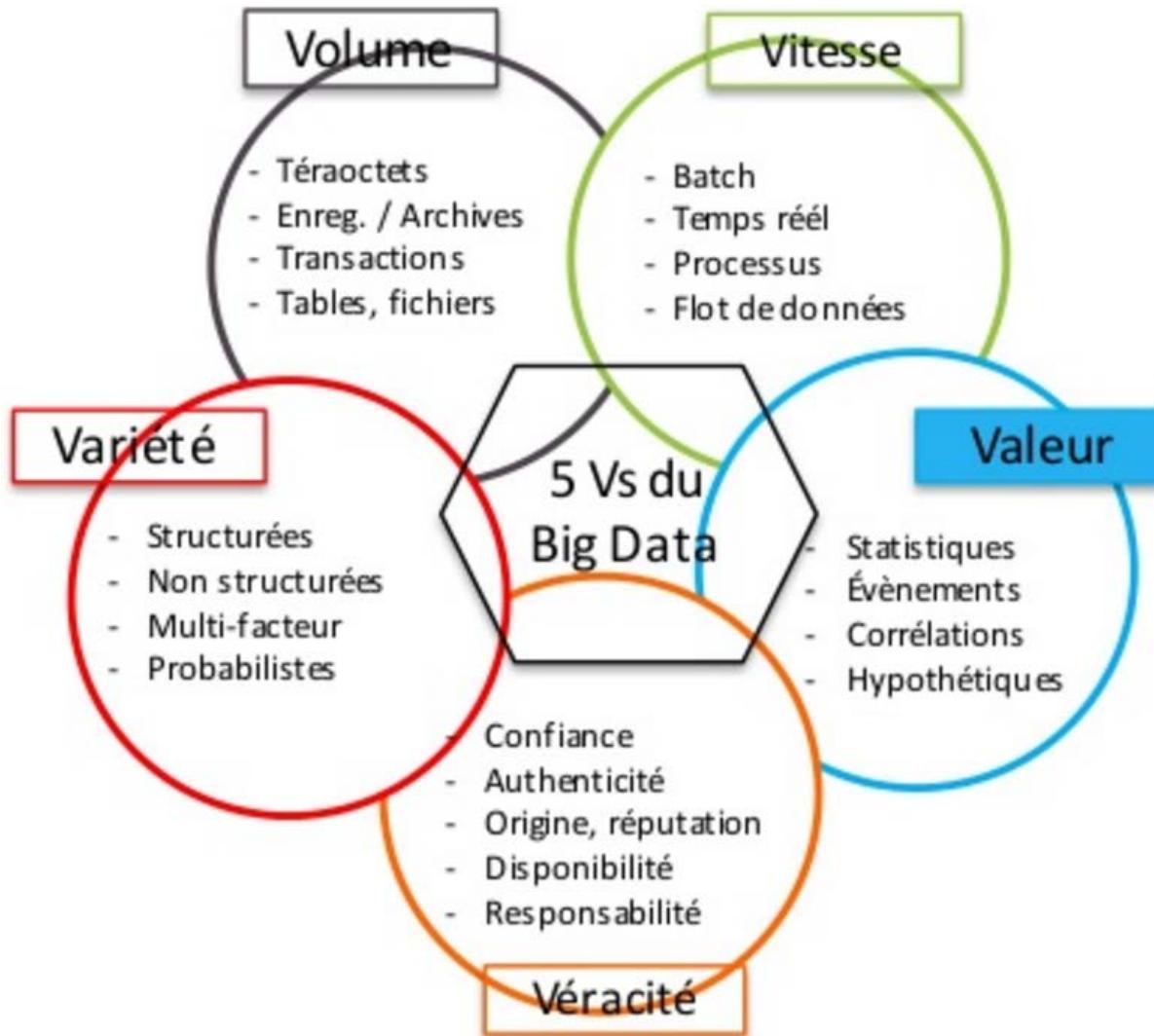
[J. van Dijck - Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology - Surveillance & Society - Vol 12 No 2 - 2014]

# Les multiples V du Big Data (2)

## □ De 3 à 6 V



# Les multiples V du Big Data - en résumé

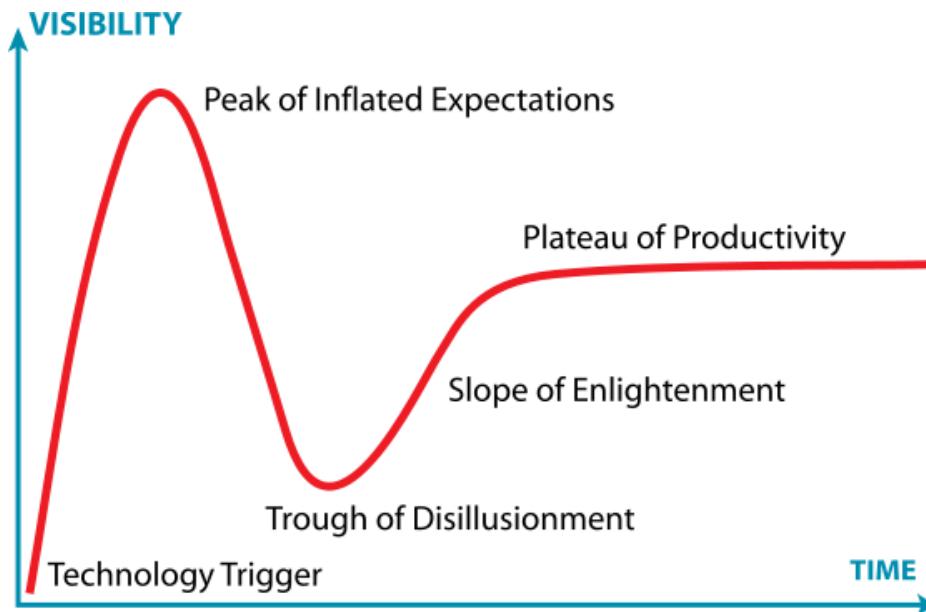


# Big Data : definition(s)

## □ Les 7 définitions du Big Data

1	The Original Big Data (3Vs)	The original type of definition is referred to Douglas Laney's Volume, Velocity and Variety or 3Vs. It has been widely cited since 2001. Many have tried to extend the number of Vs, such as 4Vs, 5Vs, 6Vs ... up to 11 Vs
2	Big Data as Technology	This type of definition is oriented by new technology development, such as MapReduce, Bulk Synchronous Parallel (BSP - Hama), Resilient Distributed Datasets (RDD, Spark), and Lambda architecture (Flink).
3	Big Data as Application	This kind of definition emphasizes different applications based on different types of Big Data. Barry Devlin defined it as application of process-mediated data, human-sourced information and machine generated data. Shaun Connolly focused on analyzing transactions, interaction and observation of data. It looks for hindsight of data.
4	Big Data as Signals	This is another type of application oriented definition but it focuses on timing rather than type of data. It looks for a foresight of data or new 'signal' pattern in dataset
5	Big Data as Opportunity	Matt Aslett: "Big Data as analyzing data that was previously ignored because of technology limitations". It highlights many potential opportunities by revisiting the collected or archived datasets when new technologies are available.
6	Big Data as Metaphor	It defines Big Data as human thinking process. It elevates BDA to the new level, which BDA is not just a type of analytics rather than the extension of human brain.
7	Big Data as New Term for Old Stuff	This definition simply means the new bottle (relabel the new term "Big Data") for old wine (Business intelligence or data mining or other traditional data analytic activities). It is one of the most cynical ways to define Big Data

# La courbe de Gartner « hype cycle »

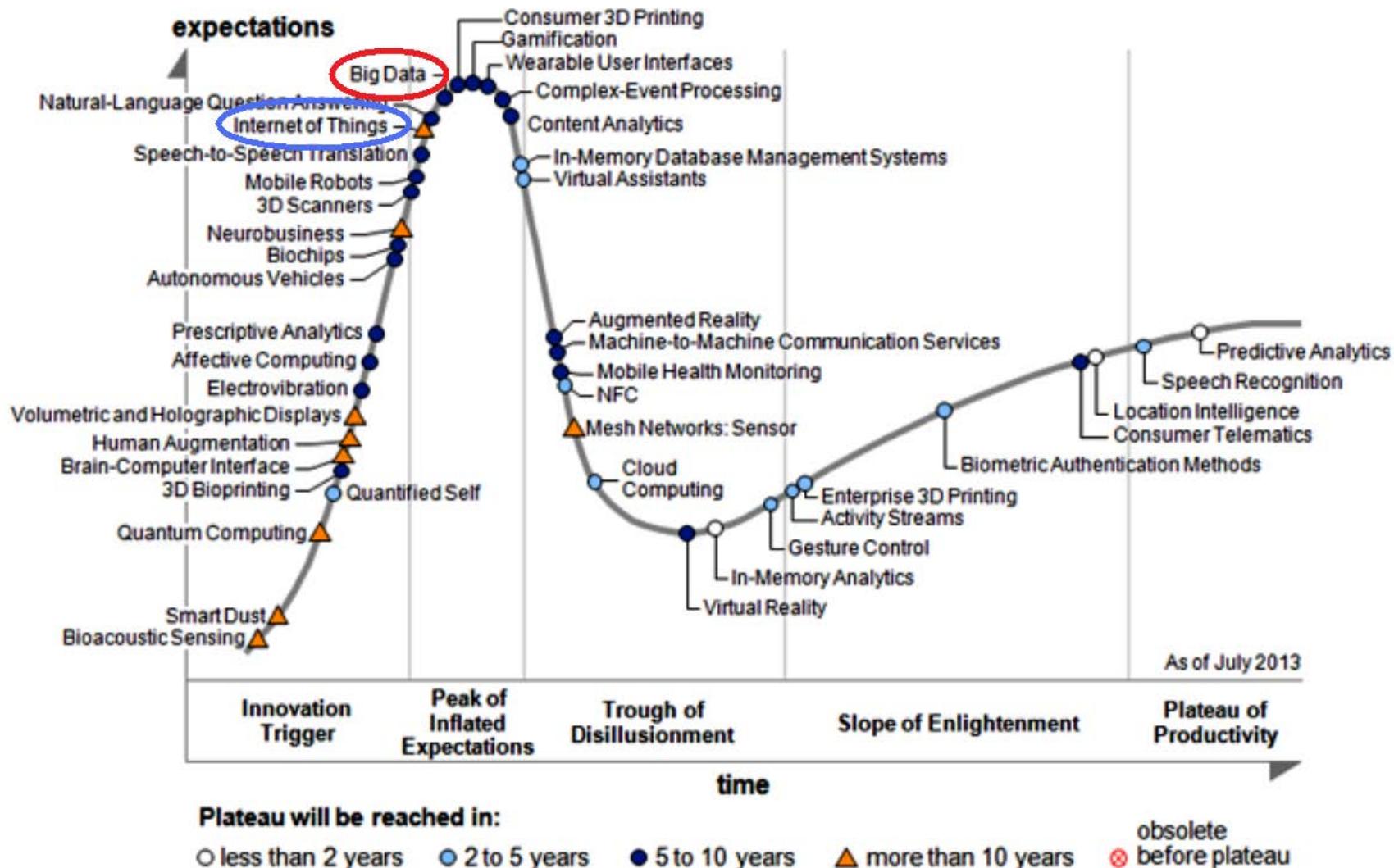


- Composée de 5 phases successives
  1. Déclenchement technologique
  2. Pic d'attente
  3. Chute de désillusions
  4. Pente de la révélation
  5. Plateau de productivité

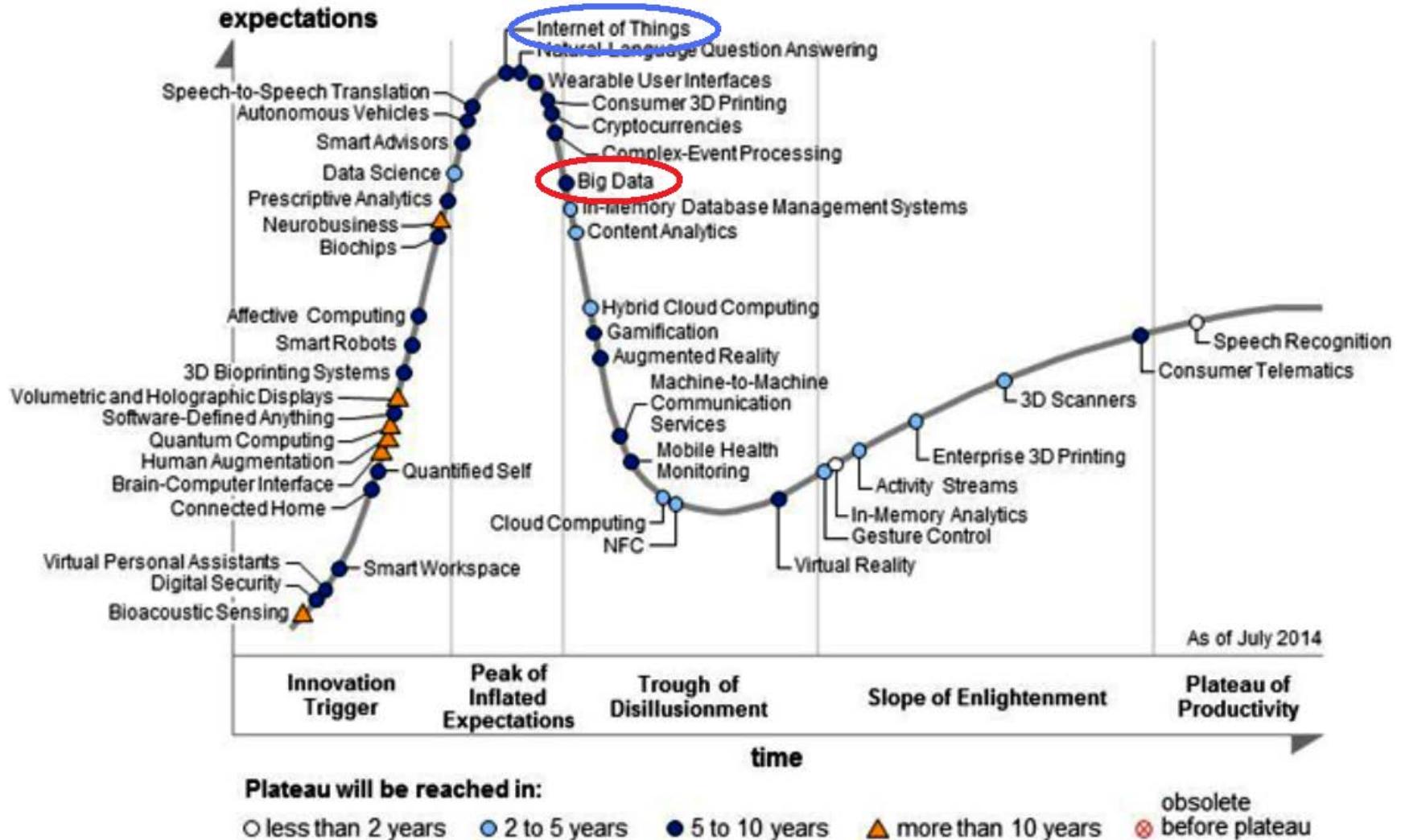
# Gartner's hype cycle - Big Data & IoT - 2012



# Gartner's hype cycle - Big Data & IoT - 2013

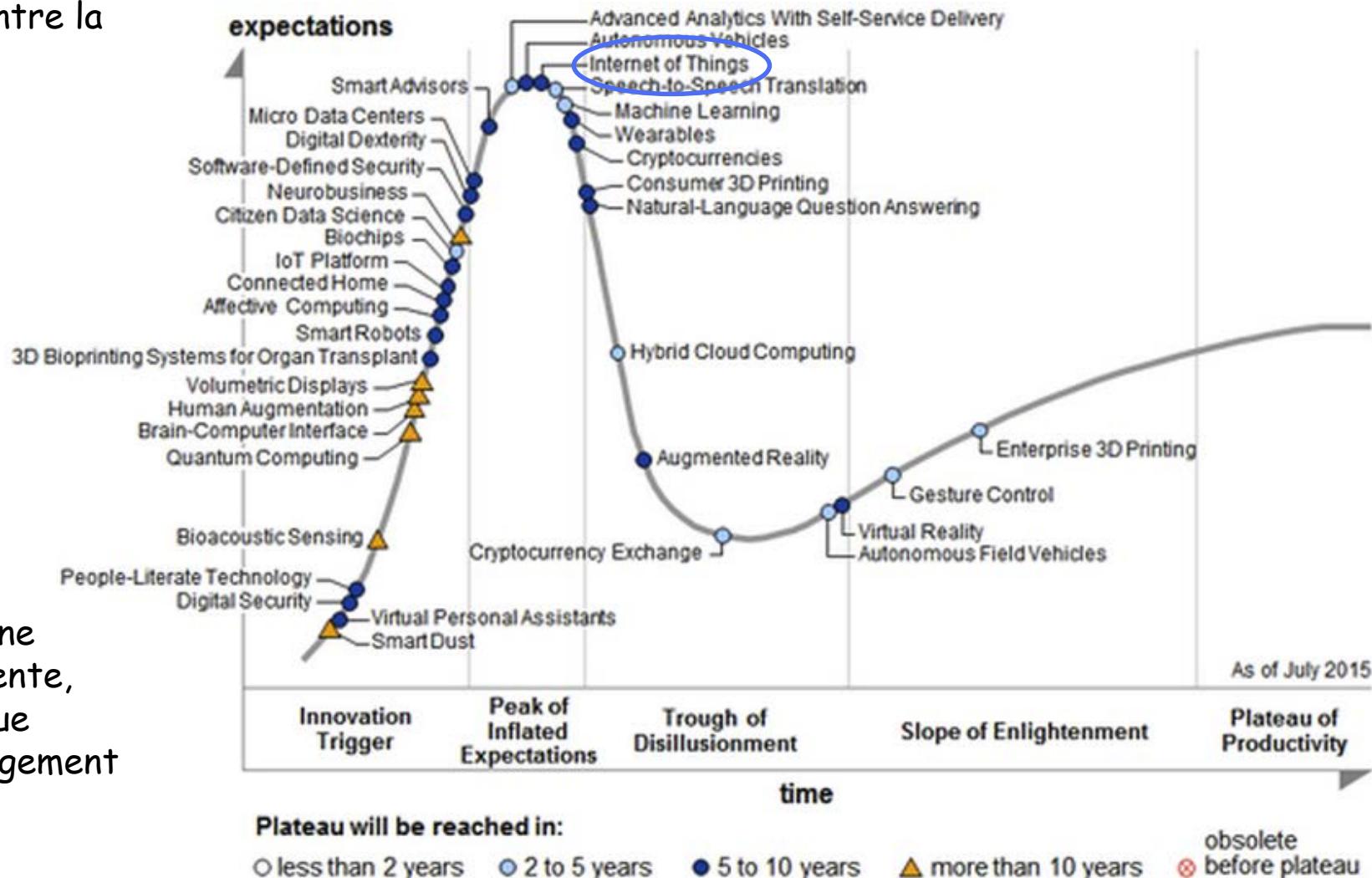


# Gartner's hype cycle - Big Data & IoT - 2014



# Gartner's hype cycle - Big Data & IoT - 2015

2 G. Perec rencontre la technologie ?



→ Ce n'est plus une techno émergente, elle est devenue une techno largement répandue

# Big Data

## Plan

### 1. Introduction

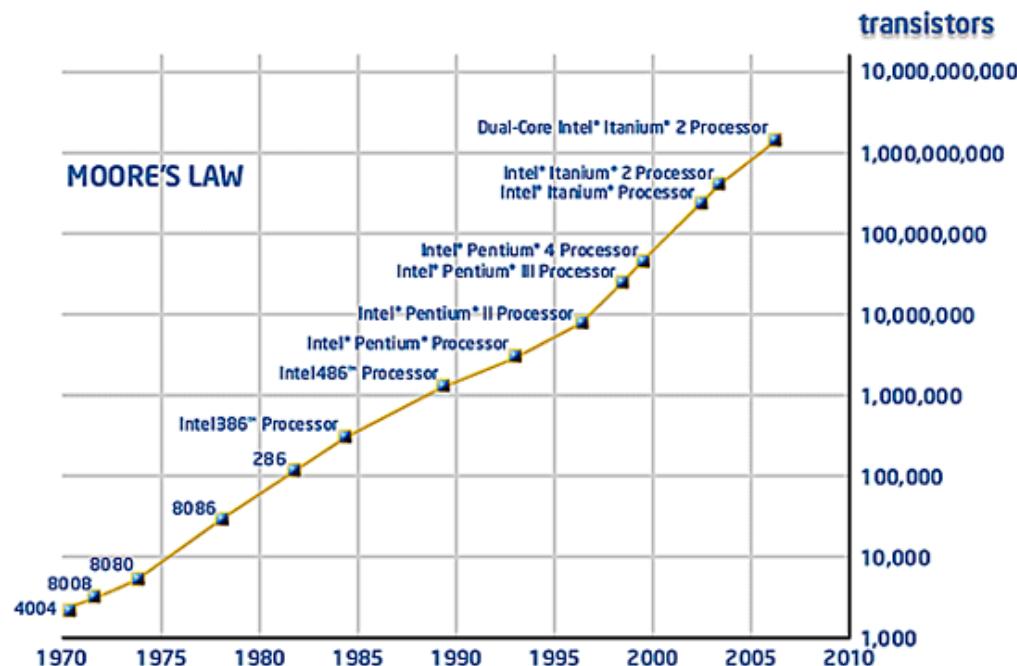
- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

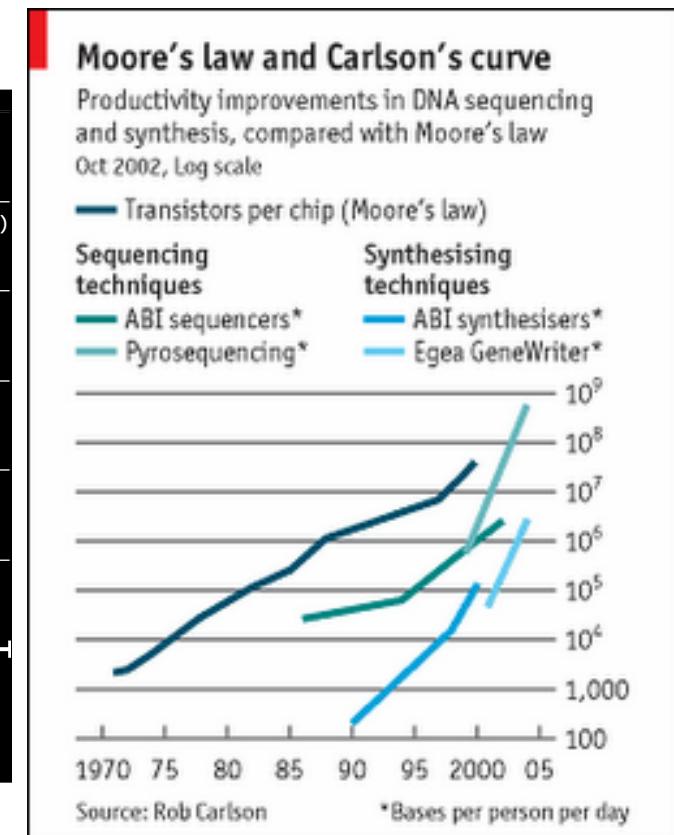
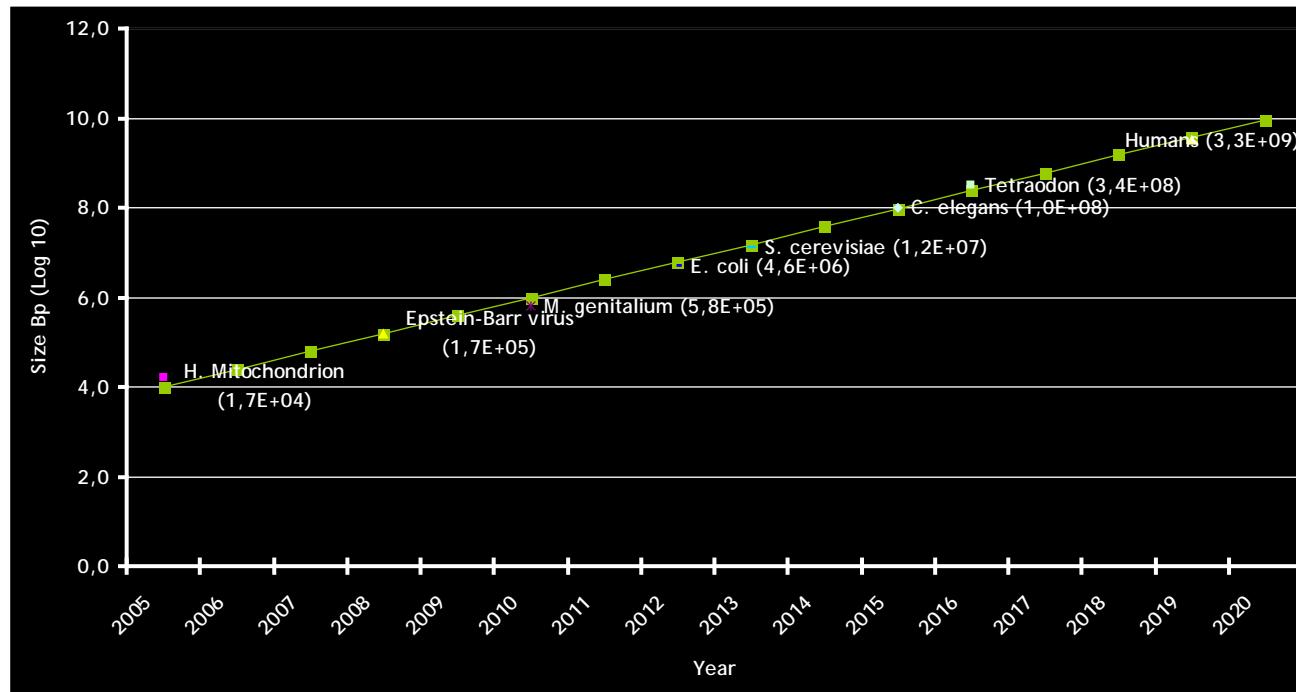
# Croissance des capacités

- Augmentation exponentielle des ressources
  - Loi de Moore (1965) : densité des transistors double tous les 24 mois



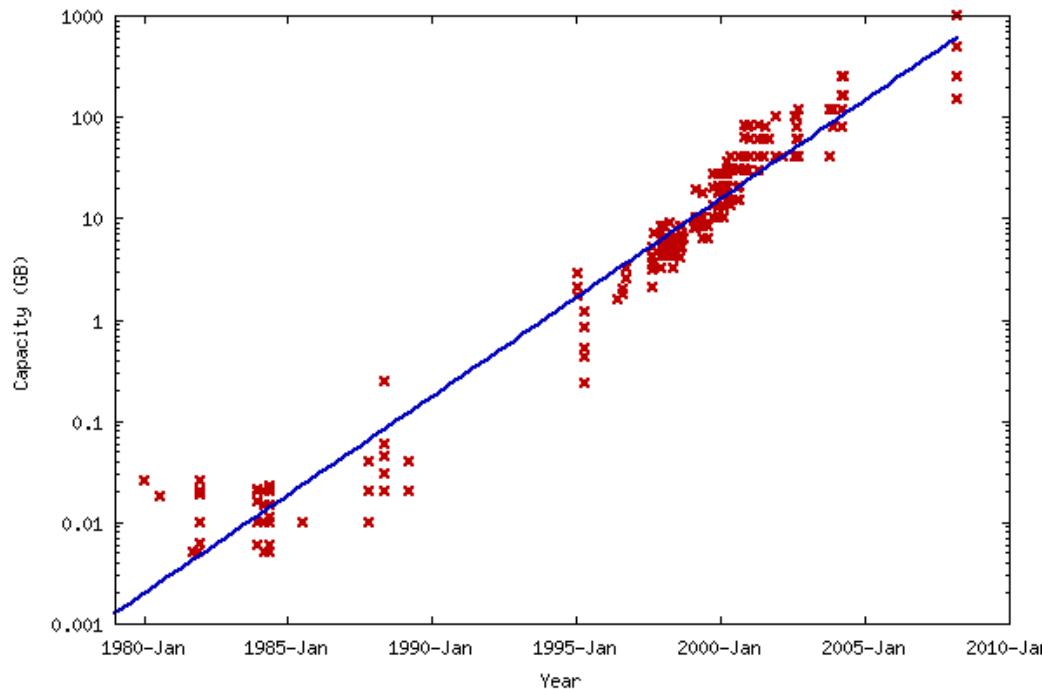
# Croissance des capacités

- Augmentation exponentielle des ressources
  - Loi de Moore (1965)
  - Carlson's curve (2002) : synthèse de l'ADN



# Croissance des capacités

- Augmentation exponentielle des ressources
  - Loi de Moore (1965)
  - Carlson's curve (2002)
  - Loi de Kryder (*Sci Am.* 2005) : capacité de stockage des disques

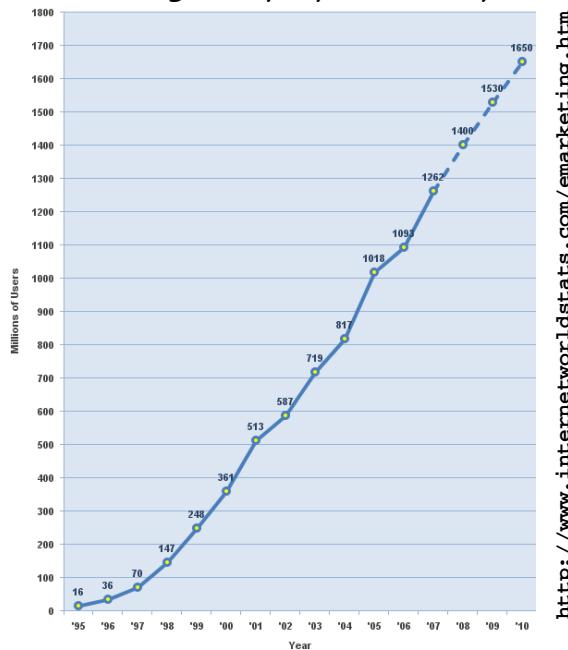


# Croissance des capacités

## □ Augmentation exponentielle des ressources

- Loi de Moore (1965)
- Carlson's curve (2002)
- Loi de Kryder (*Sci Am.* 2005)
- Loi de Metcalfe (N.Y Times - 07/1996)

(*The network is the computer : a network's value grows proportionately to the square number of its users*)



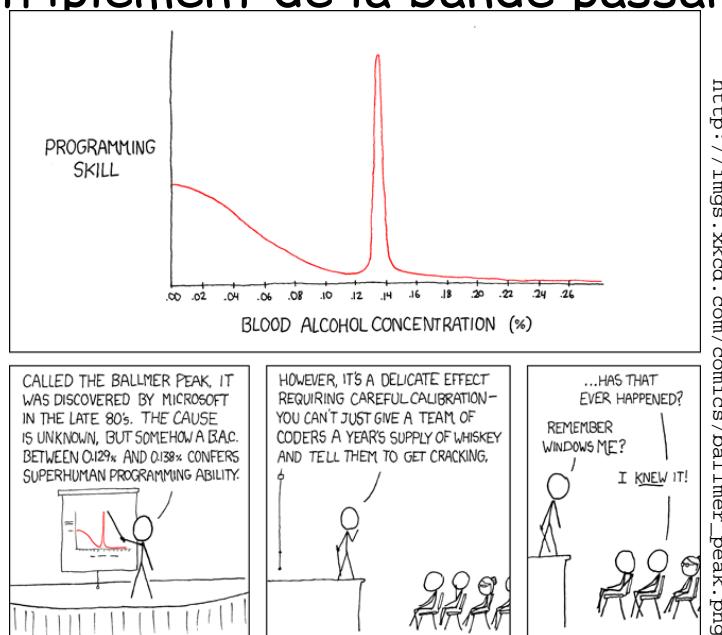
<http://www.internetworldstats.com/emarketing.htm>

# Croissance des capacités

- Augmentation exponentielle des ressources
  - Loi de Moore (1965)
  - Carlson's curve (2002)
  - Loi de Kryder (*Sci Am.* 2005)
  - Loi de Metcalfe (N.Y Times - 07/1996)
  - Loi de Gilder : triplement de la bande passante tous les 12 mois

# Croissance des capacités

- Augmentation exponentielle des ressources
  - Loi de Moore (1965)
  - Carlson's curve (2002)
  - Loi de Kryder (*Sci Am.* 2005)
  - Loi de Metcalfe (*N.Y Times* - 07/1996)
  - Loi de Gilder : triplement de la bande passante tous les 12 mois
  - Balmer's Peak

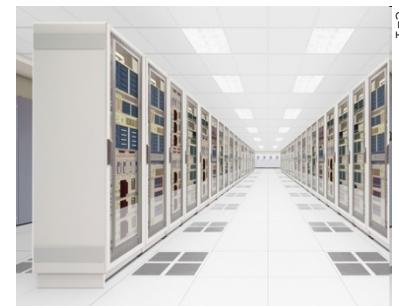


# Factorisation des ressources de calcul : le cloud

- Date des années 2000
- Rendu possible par la convergence de
  - Architectures d'application orientés services
  - Virtualisation
  - Standardisation des ressources de calcul via Internet
- ➔ Passage à l'échelle de *Software As A Service*
- Retour aux systèmes centralisés... décentralisés dans des *computing farms*
- ➔ *Cloud computing = SAAS + Grid Computing*  
*= virtualisation du logiciel*



[http://weblog.infoworld.com/tech-line/archives/cloud\\_computing/index.html?source=cloud%20computing&c=1](http://weblog.infoworld.com/tech-line/archives/cloud_computing/index.html?source=cloud%20computing&c=1)



<http://www.bdonline.co.uk/sustain/story.asp?sectioncode=662&storycode=3109389&c=1>

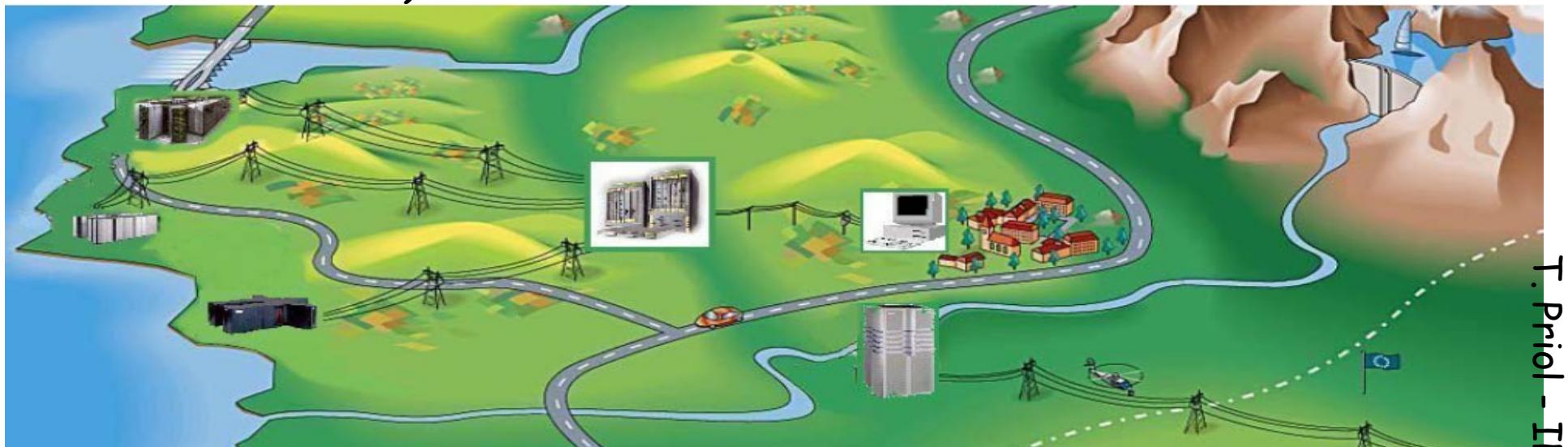
# Du grid computing (1)

- Définition du Grid-computing (Ian Foster - 07/2002)
  - Pas d'administration centralisée des ressources
  - Utilisation de standards ouverts
  - Assurance de QOS minimale
- Définition plus large par T. Priol (IRISA)
  - Une analogie avec l'énergie électrique (power grid) : Puissance de calcul = Electricité



# Du grid computing (2)

- Définition du Grid-computing (Ian Foster - 07/2002)
  - Pas d'administration centralisée des ressources
  - Utilisation de standards ouverts
  - Assurance de QOS minimale
- Définition plus large par T. Priol (IRISA)
  - Approche pour la distribution de la puissance informatique = le réseau Internet et la haute-performance (parallélisme et distribution)



# ...au Cloud Computing

- Définition du Grid-computing (Ian Foster - 07/2002)
    - Pas d'administration centralisée des ressources
    - Utilisation de standards ouverts
    - Assurance de QOS minimale
  - Définition plus large par T. Priol (IRISA)
    - Une analogie avec l'énergie électrique (power grid) : Puissance de calcul =Electricité
    - Partage coordonné de ressources dans un environnement flexible et sécurisé par une collection dynamique d'individus et d'institutions
    - Autoriser des communautés ou des organisations virtuelles à partager des ressources distribuées, dispersées géographiquement afin de poursuivre des buts communs
    - Plusieurs types de ressources  
(Processeurs, Stockage, Senseurs, Réseau, Visualisation, Logiciels, Individus, ...)
  - Très présent en informatique scientifique
    - seti@home, folding@home, grid5000...
- *Cloud Computing* : généralisation de l'utilisation de ressources (matérielles et logicielles) issues du Grid par le réseau, de façon *totalement transparente*

# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- **Volumes et production de données**
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# Volumétrie : les unités (et équivalence en temps)

Soit 1 octet  $\leftrightarrow$  1 seconde

- Ko  $10^3$  octets (17 mins)
- Mo  $10^6$  octets (12 jours)
- Go  $10^9$  octets (32 ans)
  - 1 Go : 1 semi-remorque plein de livres
  - 100 Go : un étage de bibliothèque de journaux académiques
- To  $10^{12}$  octets (33 000 ans - dernière glaciation)
  - 1 To : 50000 arbres transformés en papier et imprimés
  - 2 To : une bibliothèque de recherche universitaire
- Po  $10^{15}$  octets (moitié de la période qui nous sépare des dinosaures)
  - 2 Po : toutes les bibliothèques universitaires US
  - 20 Po : la production de disques dur en 1995
  - 200 Po : tout le matériel jamais imprimé
- Eo  $10^{18}$  octets (7 fois l'âge de la terre)
  - 2 Eo : volume total d'information produit en 1999
  - 5 Eo : tous les mots jamais prononcés par les êtres humains
- Zo  $10^{21}$  octets (2300 fois l'âge de l'univers)
- Yo  $10^{24}$  octets (2.3 millions de fois l'âge de l'univers)

# Volumétrie : les unités (et équivalence en temps)

Soit 1 octet  $\leftrightarrow$  1 seconde

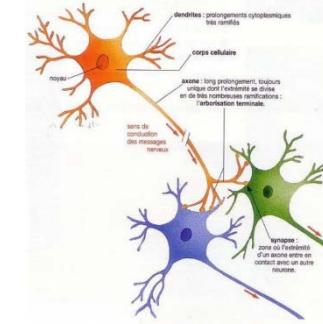
- Ko  $10^3$  octets (17 mins)
- Mo  $10^6$  octets (12 jours)
- Go  $10^9$  octets (32 ans)
  - 1 Go : 1 semi-remorque plein de livres
  - 100 Go : un étage de bibliothèque de journaux académiques
- To  $10^{12}$  octets (33 000 ans - dernière glaciation)
  - 1 To : 50000 arbres transformés en papier et imprimés
  - 2 To : une bibliothèque de recherche universitaire
- Po  $10^{15}$  octets (moitié de la période qui nous sépare des dinosaures)
  - 2 Po : toutes les bibliothèques universitaires US
  - 20 Po : la production de disques dur en 1995
  - 200 Po : tout le matériel jamais imprimé
- Eo  $10^{18}$  octets (7 fois l'âge de la terre)
  - 2 Eo : volume total d'information produit en 1999
  - 5 Eo : tous les mots jamais prononcés par les êtres humains
- Zo  $10^{21}$  octets (2300 fois l'âge de l'univers)
- Yo  $10^{24}$  octets (2.3 millions de fois l'âge de l'univers)

## Biologie

### Le cerveau humain



- 100 giga neurones
- 1 péta synapses



<http://bit.ly/2lxj4i1>

Vision : 15 mega pixels  $\times$  10 / s  
80 ans, 50 000 s/jour  $\rightarrow$  2 Eo

# Volumétrie

→ Pas de présentation sur les Big Data sans le rappel des volumes !

- Les ordres de grandeur, en Big Data :
  - Téraoctets, voire en Pétaoctets.
  - Un Petaoctet (1 Po  $\approx$  1000 Teraoctets, 1To = 1000 Gigaoctets).
  - Un Exaoctet (1 Eo  $\approx$  1000 Petaoctets), le Zettaoctets (1 Zo  $\approx$  1000 Exaoctets)
- Ordres de grandeurs (statistiques Youtube, 2014) :
  - Plus d'un milliard d'utilisateurs uniques consultent YouTube chaque mois
  - Six milliards d'heures de vidéo sur YouTube vus par mois
  - 50 % de plus qu'en 2013
  - 100 heures de vidéo mises en ligne chaque minute
  - 80 % du trafic YouTube est généré hors des États-Unis (61 pays, 61 langues)
- Twitter :
  - 58 millions de tweets échangés par jour
  - 12 téraoctets de tweets créés quotidiennement
- En 2018, il y aura 4 milliards d'utilisateurs Internet et 21 milliards d'objets connectés

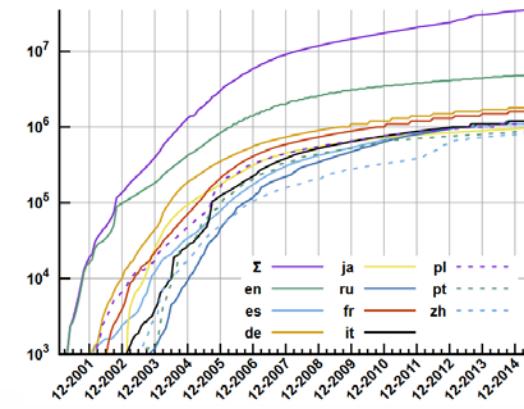
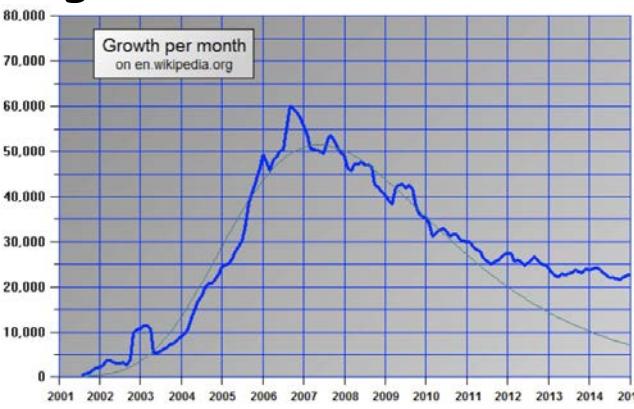
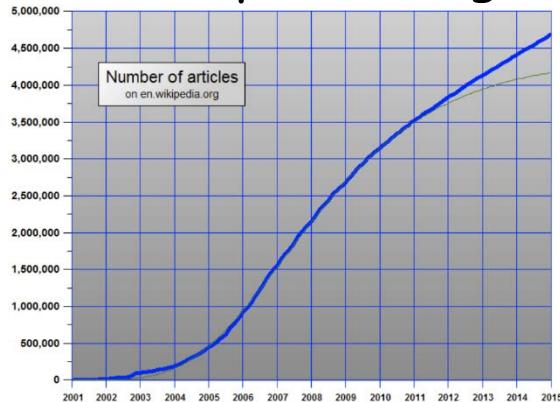
# Volumétrie (suite)

- Production de médias imprimés, film, magnétiques, optiques
  - 2002 : 5 exa-octets (Eo) de nouvelles informations en 2002
- Combien représentent 5 exa-octets ?
  - 136 To : numérisation 17 m de livres de la librairie du Congrès US
  - 5 Eo: 37000 librairies du Congrès
- 1999-2002 : 30% de croissance en informations stockées
- SMS : 5 M/jour (750 Go) ou 274 To/an
- E-mail : 400 000 To de nouvelles information/an

# Volumétrie

## □ La taille du Web (visible)

### ■ Wikipedia (langue anglaise) - [https://en.wikipedia.org/wiki/Wikipedia:Size\\_of\\_Wikipedia](https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia)

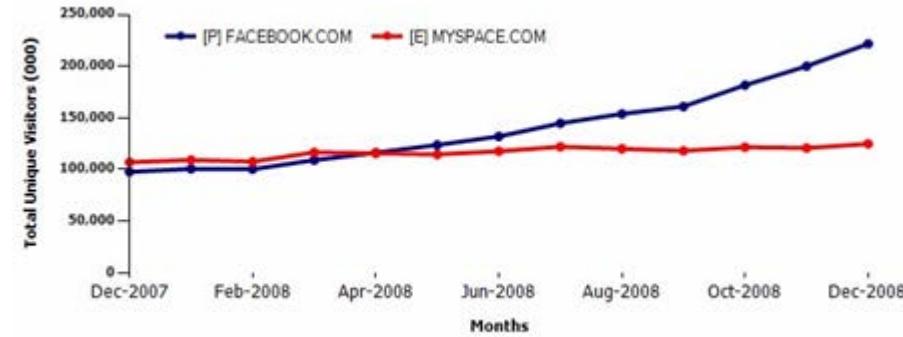


### ■ Myspace

- Plus de 50 m utilisateurs (23/1/15)

### ■ FaceBook

- Plus de 600 m utilisateurs (2011 - croissance mensuelle de 4%)



<http://www.techcrunch.com/2009/01/22/facebook-now-nearly-twice-the-size-ofmyspace-worldwide/>

## □ Volume indexé par les moteurs de recherche

### ■ 170 To d'informations

# Le web comme source de données

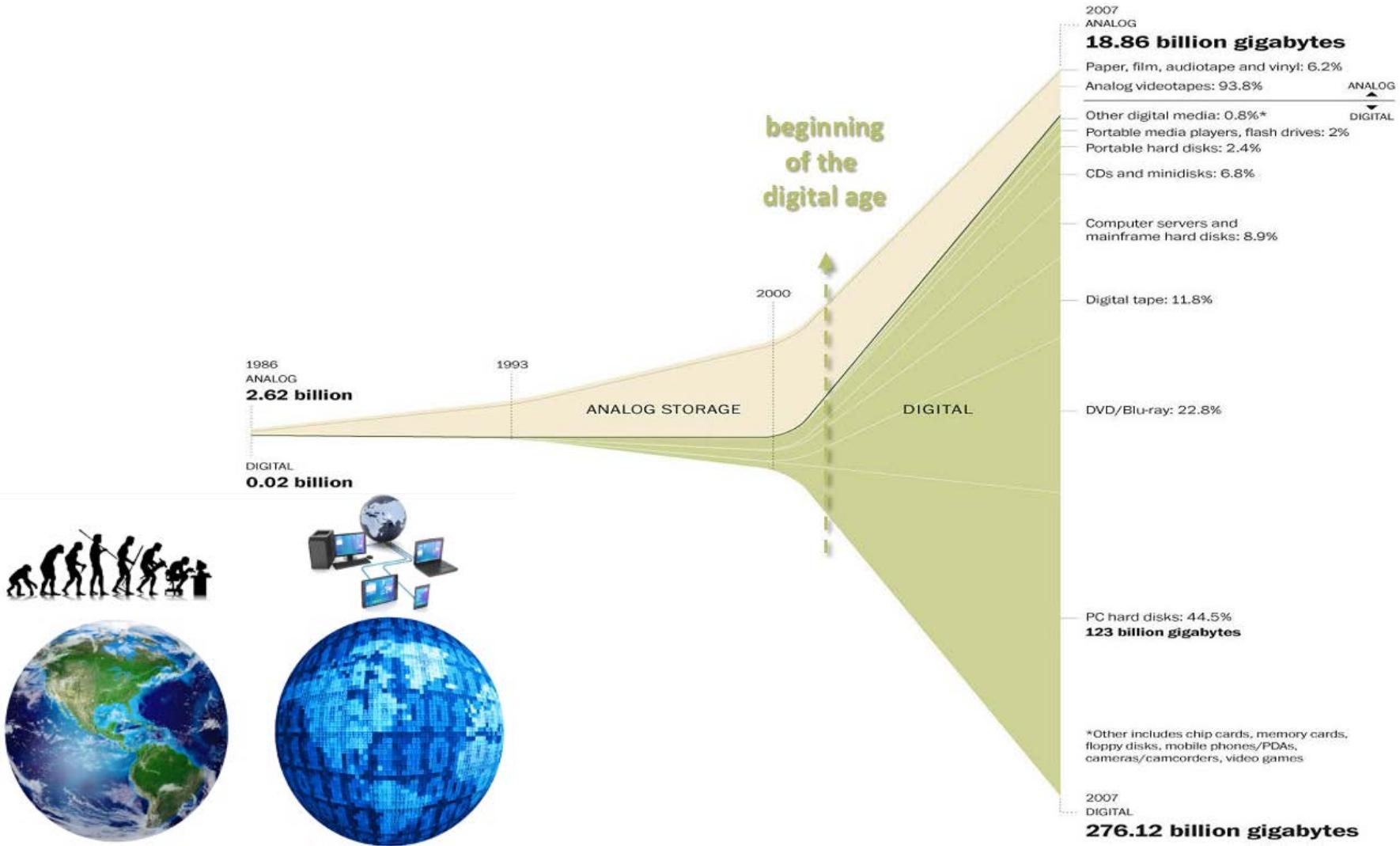
Table 2. Sixty Largest Deep Web Sites

Name	Type	URL	Web Size (GBs)
National Climatic Data Center (NOAA)	Public	<a href="http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html">http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html</a>	366,000
NASA EOSDIS	Public	<a href="http://harp.gsfc.nasa.gov/~imswww/pub/imswelcome/plain.html">http://harp.gsfc.nasa.gov/~imswww/pub/imswelcome/plain.html</a>	219,600
National Oceanographic (combined with Geophysical) Data Center (NOAA)	Public/Fee	<a href="http://www.nodc.noaa.gov/">http://www.nodc.noaa.gov/</a> , <a href="http://www.ngdc.noaa.gov/">http://www.ngdc.noaa.gov/</a>	32,940
Alexa	Public (partial)	<a href="http://www.alexa.com/">http://www.alexa.com/</a>	15,860
Right-to-Know Network (RTK Net)	Public	<a href="http://www.rtk.net/">http://www.rtk.net/</a>	14,640
MP3.com	Public	<a href="http://www.mp3.com/">http://www.mp3.com/</a>	4,300
Terraserver	Public/Fee	<a href="http://terraserver.microsoft.com/">http://terraserver.microsoft.com/</a>	4,270
HEASARC (High Energy Astrophysics Science Archive Research Center)	Public	<a href="http://heasarc.gsfc.nasa.gov/W3Browse/">http://heasarc.gsfc.nasa.gov/W3Browse/</a>	2,562
US PTO - Trademarks + Patents	Public	<a href="http://www.uspto.gov/tmdb/">http://www.uspto.gov/tmdb/</a> , <a href="http://www.uspto.gov/patft/">http://www.uspto.gov/patft/</a>	2,440
Informedia (Carnegie Mellon Univ.)	Public (not yet)	<a href="http://www.informedia.cs.cmu.edu/">http://www.informedia.cs.cmu.edu/</a>	1,830
Alexandria Digital Library	Public	<a href="http://www.alexandria.ucsb.edu/adl.html">http://www.alexandria.ucsb.edu/adl.html</a>	1,220
JSTOR Project	Limited	<a href="http://www.jstor.org/">http://www.jstor.org/</a>	1,220
10K Search Wizard	Public	<a href="http://www.tenkwizard.com/">http://www.tenkwizard.com/</a>	769
UC Berkeley Digital Library Project	Public	<a href="http://elib.cs.berkeley.edu/">http://elib.cs.berkeley.edu/</a>	766
SEC Edgar	Public	<a href="http://www.sec.gov/edgarhp.htm">http://www.sec.gov/edgarhp.htm</a>	610
US Census	Public	<a href="http://factfinder.census.gov">http://factfinder.census.gov</a>	610
NCI CancerNet Database	Public	<a href="http://cancernet.nci.nih.gov/">http://cancernet.nci.nih.gov/</a>	488
Amazon.com	Public	<a href="http://www.amazon.com/">http://www.amazon.com/</a>	461
IBM Patent Center	Public/Private	<a href="http://www.patents.ibm.com/boolquery">http://www.patents.ibm.com/boolquery</a>	345
NASA Image Exchange	Public	<a href="http://nix.nasa.gov/">http://nix.nasa.gov/</a>	337
InfoUSA.com	Public/Private	<a href="http://www.abii.com/">http://www.abii.com/</a>	195
Betterwhois (many similar)	Public	<a href="http://betterwhois.com/">http://betterwhois.com/</a>	152
GPO Access	Public	<a href="http://www.access.gpo.gov/">http://www.access.gpo.gov/</a>	146
Adobe PDF Search	Public	<a href="http://searchpdf.adobe.com/">http://searchpdf.adobe.com/</a>	143
Internet Auction List	Public	<a href="http://www.internetauctionlist.com/search_products.html">http://www.internetauctionlist.com/search_products.html</a>	130
Commerce, Inc.	Public	<a href="http://search.commerceinc.com/">http://search.commerceinc.com/</a>	122
Library of Congress Online Catalog	Public	<a href="http://catalog.loc.gov/">http://catalog.loc.gov/</a>	116
Sunsite Europe	Public	<a href="http://src.doc.ic.ac.uk/">http://src.doc.ic.ac.uk/</a>	98
Uncover Periodical DB	Public/Fee	<a href="http://uncweb.carl.org/">http://uncweb.carl.org/</a>	97
Astronomer's Bazaar	Public	<a href="http://cdsweb.u-strasbg.fr/Cats.html">http://cdsweb.u-strasbg.fr/Cats.html</a>	94
eBay.com	Public	<a href="http://www.ebay.com/">http://www.ebay.com/</a>	82
REALTOR.com Real Estate Search	Public	<a href="http://www.realtor.com/">http://www.realtor.com/</a>	60
Federal Express	Public (if shipper)	<a href="http://www.fedex.com/">http://www.fedex.com/</a>	53
Integrum	Public/Private	<a href="http://www.integrumworld.com/eng_test/index.html">http://www.integrumworld.com/eng_test/index.html</a>	49
NIH PubMed	Public	<a href="http://www.ncbi.nlm.nih.gov/PubMed/">http://www.ncbi.nlm.nih.gov/PubMed/</a>	41

<http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104>

Visual Woman (NIH)	Public	<a href="http://www.nlm.nih.gov/research/visible/visible_human.html">http://www.nlm.nih.gov/research/visible/visible_human.html</a>	40
AutoTrader.com	Public	<a href="http://www.autoconnect.com/index.jtmpl/?LNX=M1DJAROSTEXT">http://www.autoconnect.com/index.jtmpl/?LNX=M1DJAROSTEXT</a>	39
UPS	Public (if shipper)	<a href="http://www.ups.com/">http://www.ups.com/</a>	33
NIH GenBank	Public	<a href="http://www.ncbi.nlm.nih.gov/Genbank/index.html">http://www.ncbi.nlm.nih.gov/Genbank/index.html</a>	31
AustLII (Australasian Legal Information Institute)	Public	<a href="http://www.austlii.edu.au/austlii/">http://www.austlii.edu.au/austlii/</a>	24
Digital Library Program (UVA)	Public	<a href="http://www.lva.lib.va.us/">http://www.lva.lib.va.us/</a>	21
<b>Subtotal Public and Mixed Sources</b>			<b>673,035</b>
DBT Online	Fee	<a href="http://www.dbtonline.com/">http://www.dbtonline.com/</a>	30,500
Lexis-Nexis	Fee	<a href="http://www.lexis-nexis.com/lnc/">http://www.lexis-nexis.com/lnc/</a>	12,200
Dialog	Fee	<a href="http://www.dialog.com/">http://www.dialog.com/</a>	10,980
Genealogy - ancestry.com	Fee	<a href="http://www.ancestry.com/">http://www.ancestry.com/</a>	6,500
ProQuest Direct (incl. Digital Vault)	Fee	<a href="http://www.umi.com">http://www.umi.com</a>	3,172
Dun & Bradstreet	Fee	<a href="http://www.dnb.com">http://www.dnb.com</a>	3,113
Westlaw	Fee	<a href="http://www.westlaw.com/">http://www.westlaw.com/</a>	2,684
Dow Jones News Retrieval	Fee	<a href="http://dowjones.wsj.com/p/main.html">http://dowjones.wsj.com/p/main.html</a>	2,684
infoUSA	Fee/Public	<a href="http://www.infousa.com/">http://www.infousa.com/</a>	1,584
Elsevier Press	Fee	<a href="http://www.elsevier.com">http://www.elsevier.com</a>	570
EBSCO	Fee	<a href="http://www.ebsco.com">http://www.ebsco.com</a>	481
Springer-Verlag	Fee	<a href="http://link.springer.de/">http://link.springer.de/</a>	221
OVID Technologies	Fee	<a href="http://www.ovid.com">http://www.ovid.com</a>	191
Investext	Fee	<a href="http://www.investext.com/">http://www.investext.com/</a>	157
Blackwell Science	Fee	<a href="http://www.blackwell-science.com">http://www.blackwell-science.com</a>	146
GenServ	Fee	<a href="http://gs01.genserv.com/gs/bcc.htm">http://gs01.genserv.com/gs/bcc.htm</a>	106
Academic Press IDEAL	Fee	<a href="http://www.idealibrary.com">http://www.idealibrary.com</a>	104
Tradecompass	Fee	<a href="http://www.tradecompass.com/">http://www.tradecompass.com/</a>	61
INSPEC	Fee	<a href="http://www.iee.org.uk/publish/inspec/online/online.html">http://www.iee.org.uk/publish/inspec/online/online.html</a>	16
<b>Subtotal Fee-Based Sources</b>			<b>75,469</b>
<b>TOTAL</b>			<b>748,504</b>

# Big Data et volume des données



<https://www.youtube.com/watch?v=J0bp2kUh9hw>

# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- **Quelques projets emblématiques**
- Technologie et écosystème
- Les entreprises

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# Un projet Small Data (1)

- De <http://www.ovh.com/fr/a1136.interview-github-octave-klabo-ovh>
- O. Klabo (OVH) interdit l'usage de GitHub pour ses développeurs



@MrTAZ42 @gierschv la dernière fois qu'on a partagé un patch, on nous a débauché le gars. depuis j'ai interdit de publier les diff's.

[Répondre](#) [Retweeter](#) [Favori](#) [Plus](#)

# Un projet Small Data (2)

- De <http://www.ovh.com/fr/a1136.interview-github-octave-klabo-ovh>
- O. Klabo (OVH) interdit l'usage de GitHub pour ses développeurs



@PoolpOrg @foo\_ @MrTAZ42 @gierschv  
les sites comme github permettent de répertorier, faire les stats, les actions sur la communauté de dev.

◀ Répondre ↗ Retweeter ★ Favori ⚡ Plus

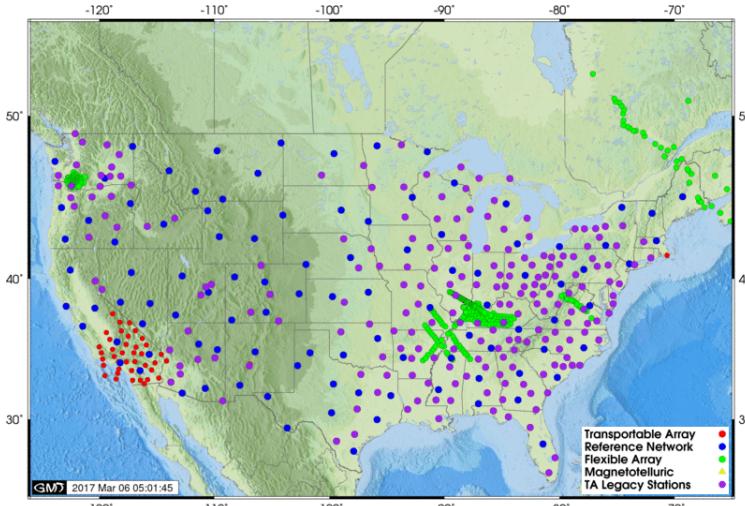
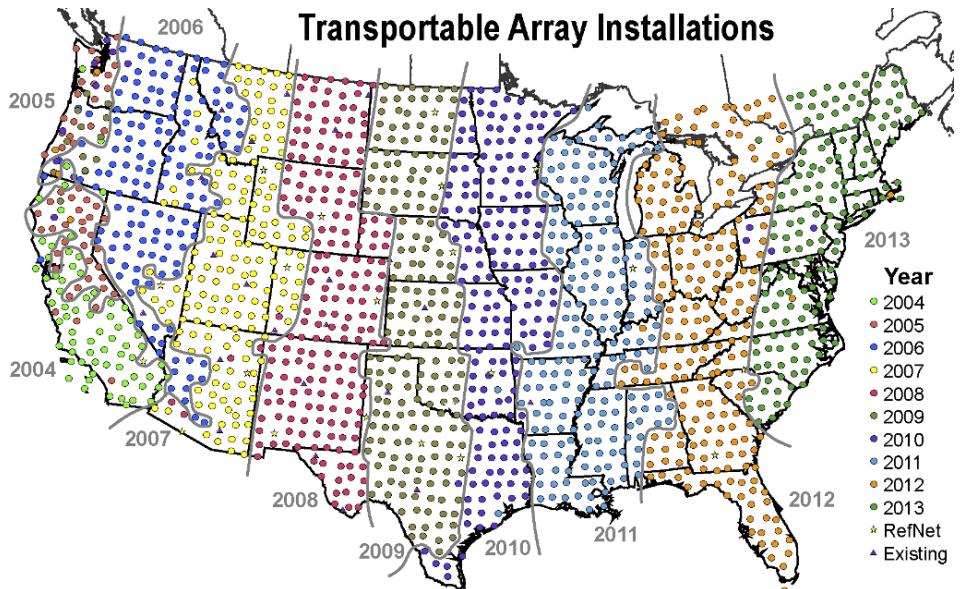
- « Si un cabinet de recrutement est capable de prendre contact avec un développeur, cela signifie que n'importe quelle entreprise peut recueillir des données sur tous les développeurs d'une entreprise concurrente : combien de développeurs a-t-elle ? Quelle est la spécialité de chacun ? Combien de contributions chaque développeur fait-il par semaine ? Sur quoi travaille-t-il en ce moment ? Beaucoup d'informations très intéressantes que nous ne souhaitons pas partager avec nos concurrents. »

# Des projets de Big Data (1)

## □ Sciences de la terre

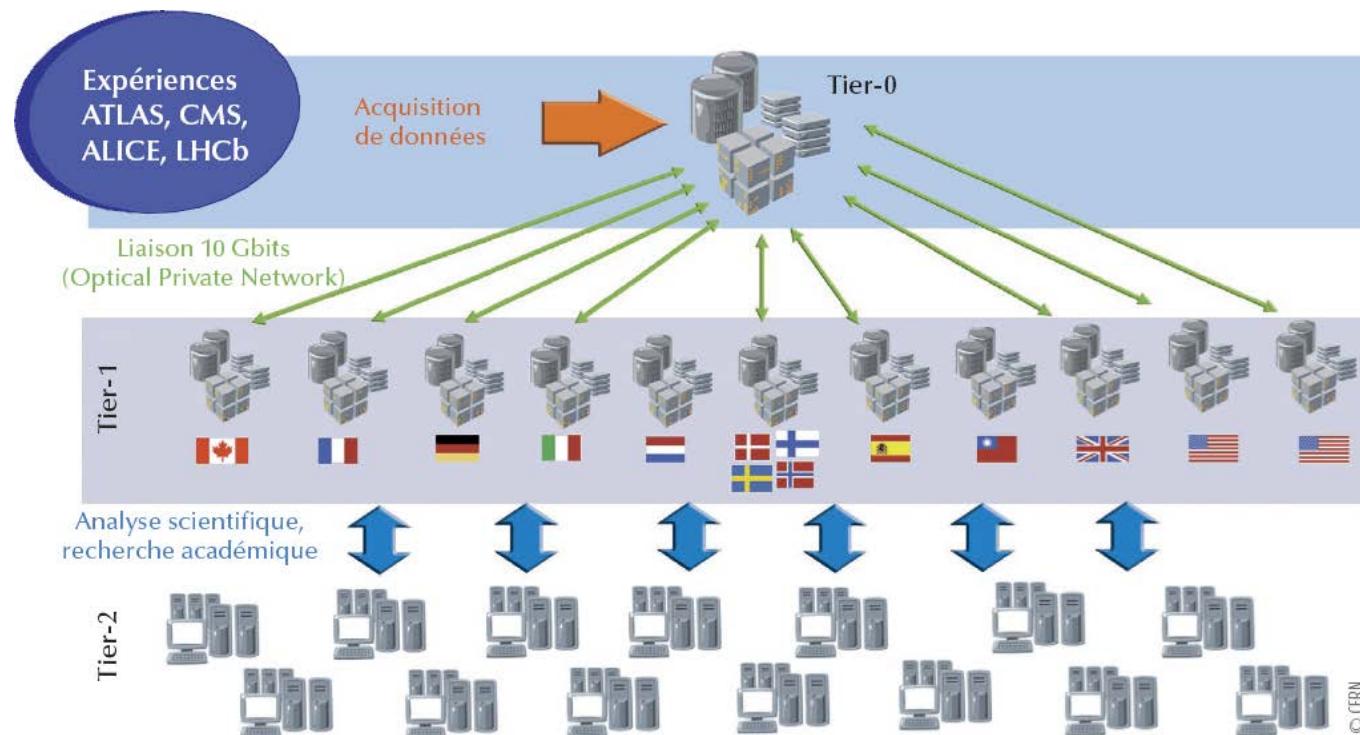
USArray : a continental-scale seismic observatory : 400 sismographes et capteurs atmosphériques

## □ État au 06/03/2017



# Des projets de Big Data (2)

## □ CERN, LHC et Boson de Higgs



- Accélérateur de particules produisant des collisions proton-proton. Environ une centaine de milliards de particules sont accélérées quasiment à la vitesse de la lumière et entrent en collision toutes les secondes. En théorie, un seul boson de Higgs est produit tous les 10 milliards de collisions.
- Lors d'une expérience, les capteurs produisent 1Go/s ( $10^9$  octets) de données soit environ 30 Po ( $10^{15}$ ) par année.
- Le traitement des données nécessite le recours à une architecture de type Grid en 3 niveaux.

# Des projets de Big Data (3)

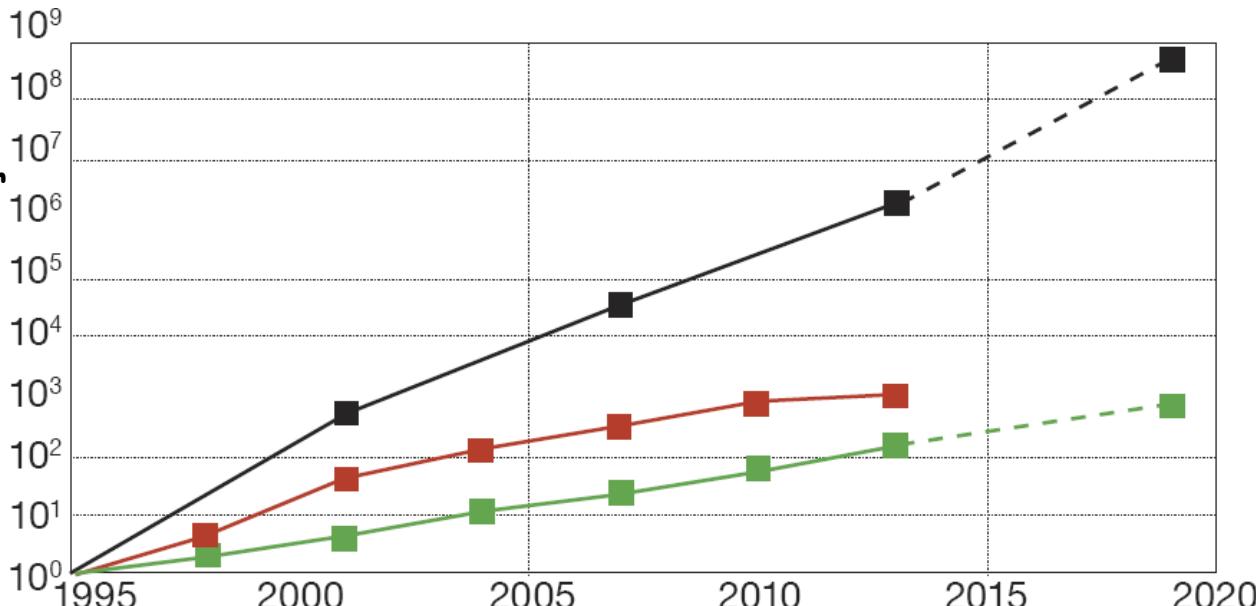
## □ Climat

« Le climat peut se définir comme le comportement statistique du temps qu'il fait, et son étude nécessite des séries de données les plus complètes possible. »

[Jean-Louis Dufresne et Sébastien Denvil ]

En 20 ans [1995-2013],

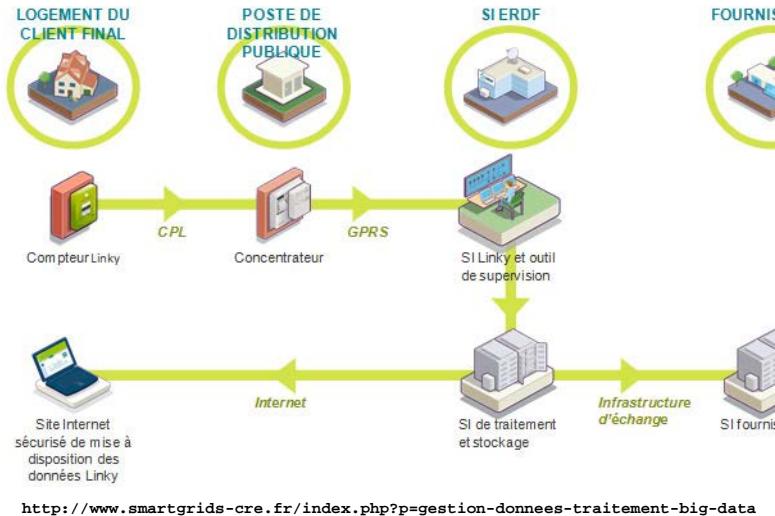
- volume de données  $\times 10^9$ ,
- densité de stockage  $\times 10^3$  pour les disques et
- $\times 150$  pour les bandes



- noir : volume de données générées par les projets coordonnés de simulations climatiques,
- rouge : quantité de données qui peut être stockée par unité de surface pour les disques et
- vert : les bandes magnétiques.

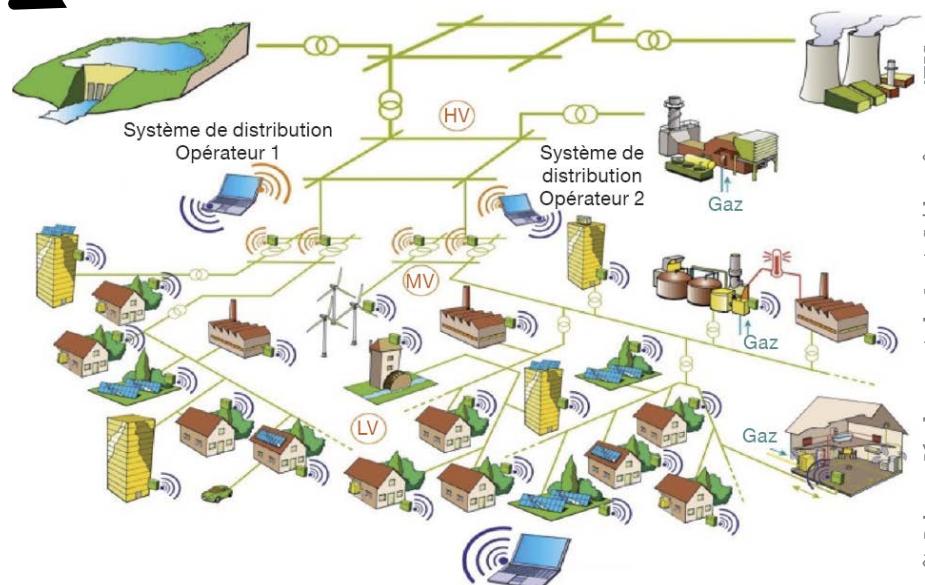
# Des projets de Big Data (4)

## □ Du « compteur intelligent »...



→ Linky : 35 millions de clients

...aux smart-grids



# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- **Technologie et écosystème**
- Les entreprises

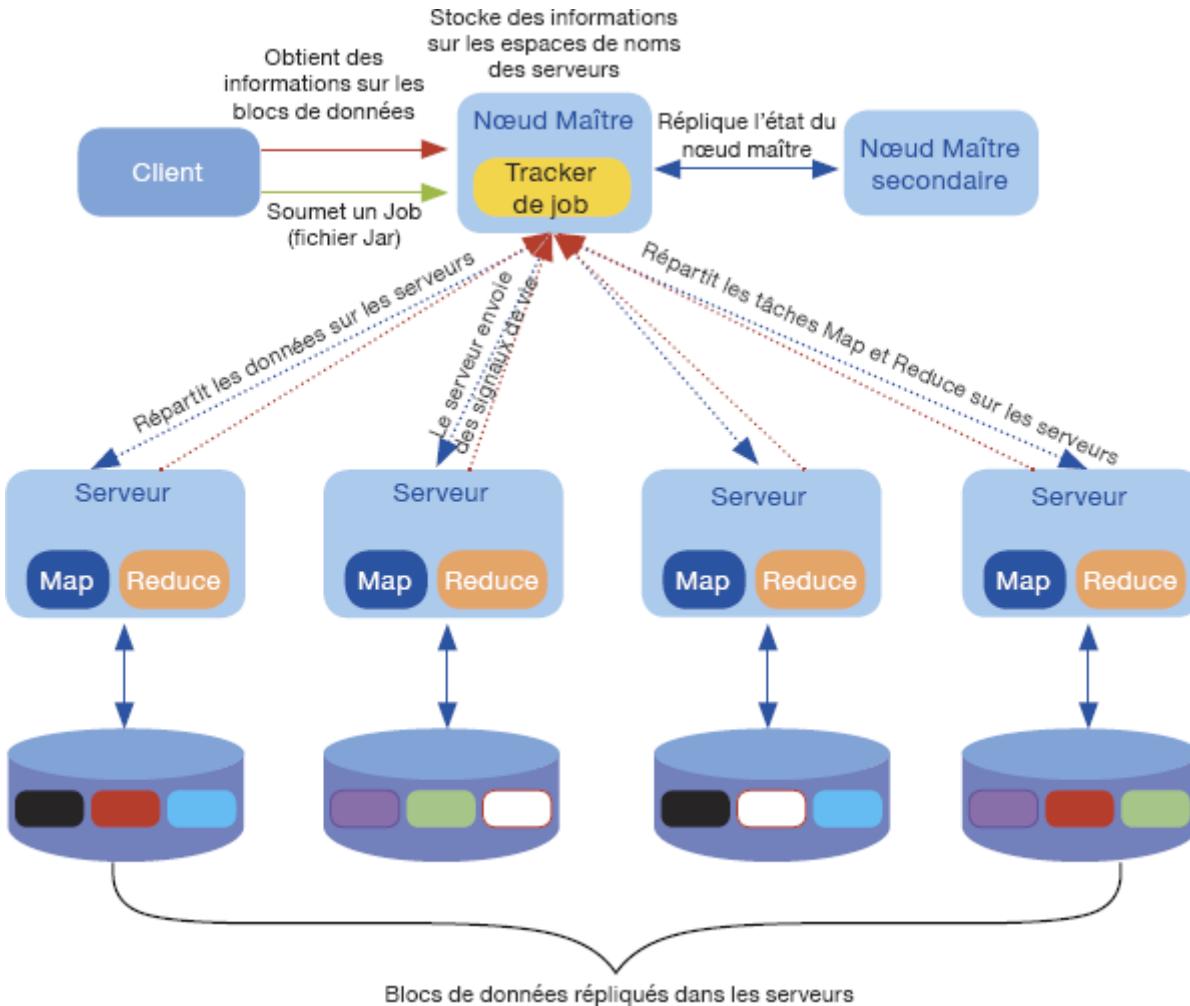
"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

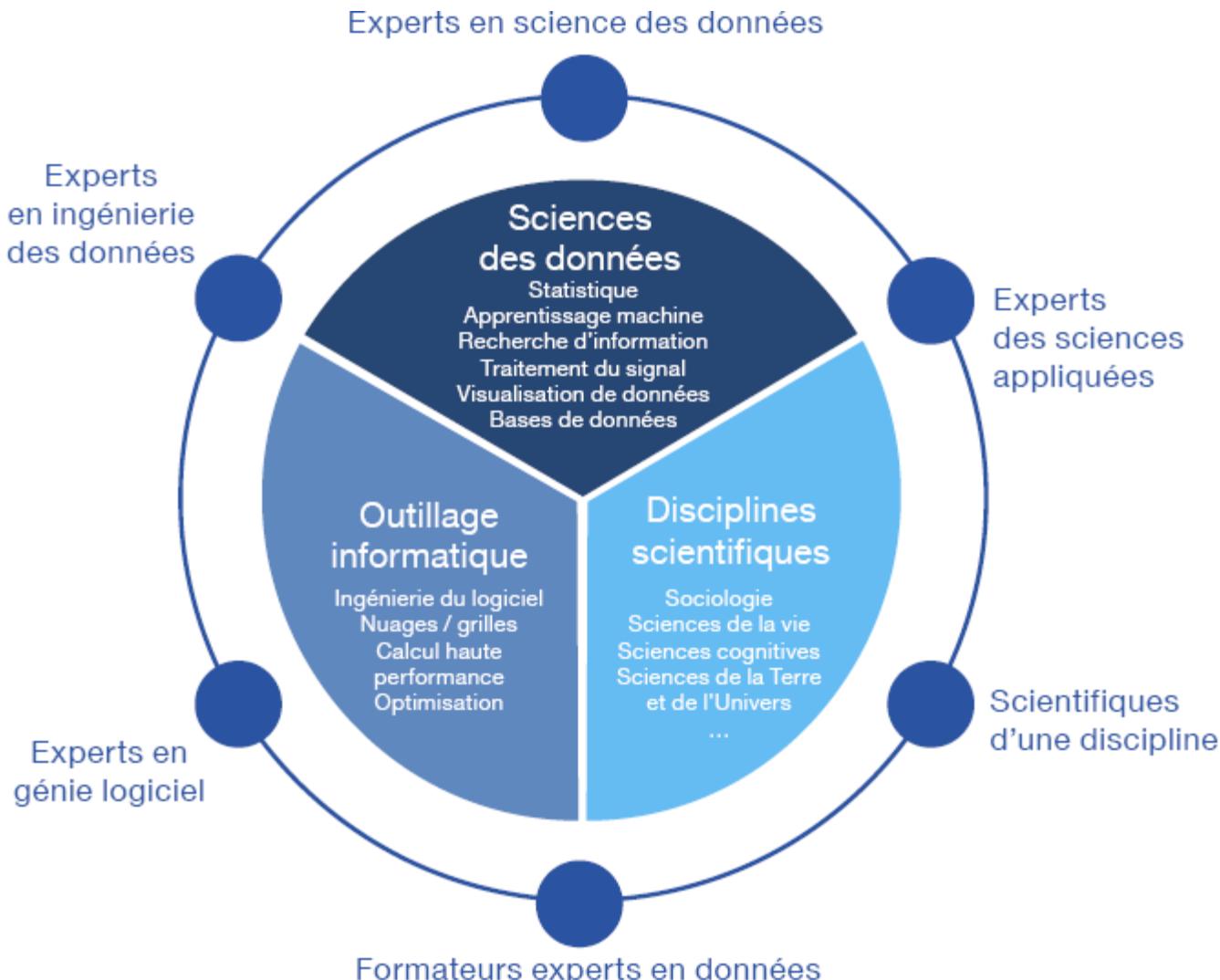
# Hadoop

## □ Architecture de référence

→ cours Big Data 2



# L'écosystème des Big Data



# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- [Les entreprises](#)

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

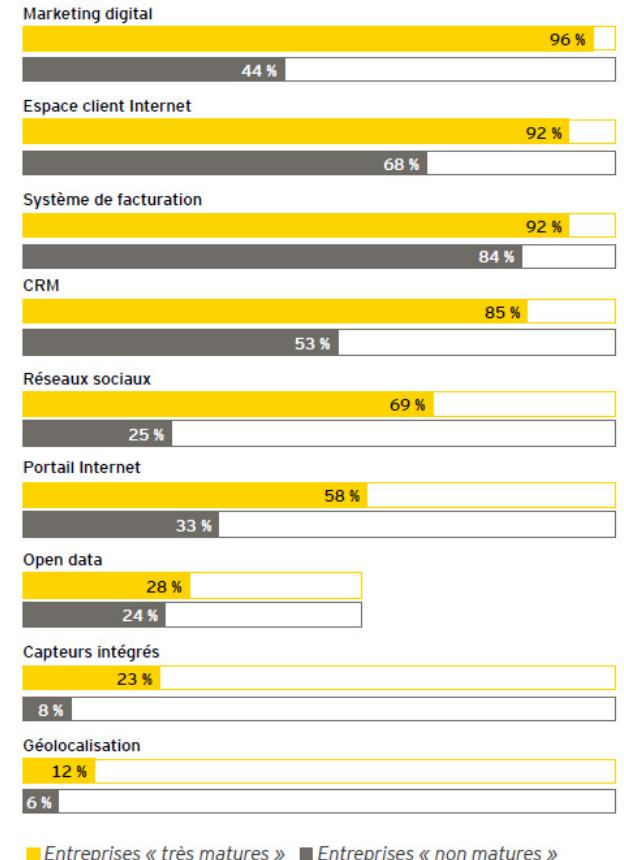
*Information Rules, C. Shapiro, Hal R. Varian*

# Big Data : les 10 freins au développement (EY)

- Selon EY (ex-Ernst&Young), analyse de 2014

[<http://www.ey.com/fr/fr/services/advisory/ey-etude-big-data-2014-10-freins-identifies>]

## 1. La collecte de la data encore largement limitée aux canaux traditionnels



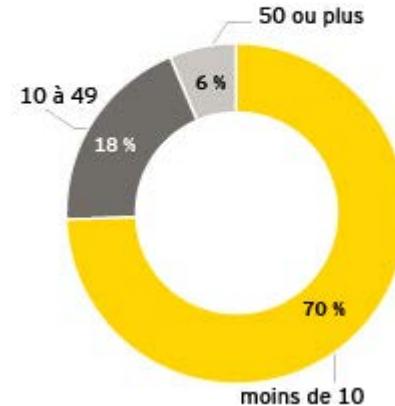
# Big Data : les 10 freins au développement (EY)

## □ Selon EY (ex-Ernst&Young), analyse de 2014

[<http://www.ey.com/fr/fr/services/advisory/ey-etude-big-data-2014-10-freins-identifies>]

1. La collecte de la data encore largement limitée aux canaux traditionnels
2. Les données non structurées, le maillon faible de l'analyse  
(45 % des entreprises interrogées collectent des données textes non structurées)
3. Un manque de compétences analytiques

Nombre de personnes dédiées à l'exploitation de la data  
Total panel : 152 entreprises



# Big Data : les 10 freins au développement (EY)

## □ Selon EY (ex-Ernst&Young), analyse de 2014

[<http://www.ey.com/fr/fr/services/advisory/ey-etude-big-data-2014-10-freins-identifies>]

1. La collecte de la data encore largement limitée aux canaux traditionnels
2. Les données non structurées, le maillon faible de l'analyse
3. Un manque de compétences analytiques
4. Une carence des outils de traitement des données
5. L'analyse de la data encore (trop) peu orientée vers le prédictif et le temps réel

[Seules 10 % des entreprises interrogées exploitent leurs données clients à des fins prédictives et 5 % d'entre elles le font pour optimiser les process techniques permettant d'accroître rapidité d'exécution et augmentation des capacités de stockage]

# Big Data : les 10 freins au développement (EY)

## □ Selon EY (ex-Ernst&Young), analyse de 2014

[<http://www.ey.com/fr/fr/services/advisory/ey-etude-big-data-2014-10-freins-identifies>]

1. La collecte de la data encore largement limitée aux canaux traditionnels
2. Les données non structurées, le maillon faible de l'analyse
3. Un manque de compétences analytiques
4. Une carence des outils de traitement des données
5. L'analyse de la data encore (trop) peu orientée vers le prédictif et le temps réel
6. Le manque de transversalité dans la gestion des projets (Big) data

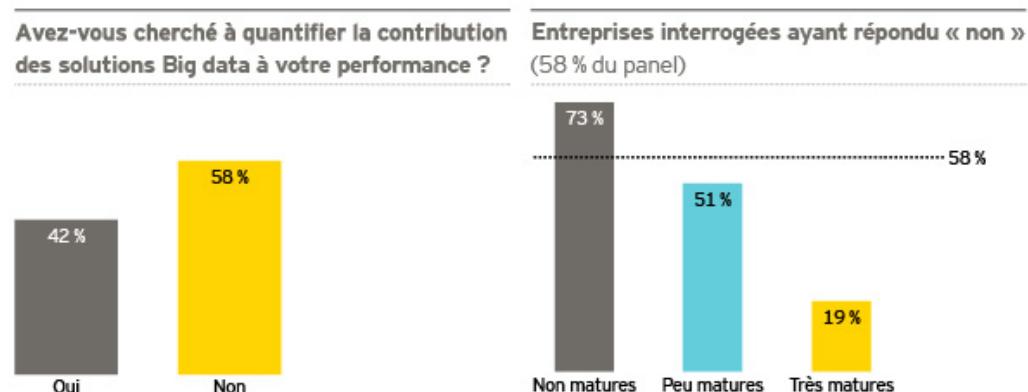
[Chaque métier ayant pour habitude d'utiliser et de transformer les données issues de ses bases de données pour répondre à ses propres enjeux métiers ou objectifs, le capital data ne peut pas circuler dans l'entreprise, ce qui explique une absence de vision unifiée]

# Big Data : les 10 freins au développement (EY)

## □ Selon EY (ex-Ernst&Young), analyse de 2014

[<http://www.ey.com/fr/fr/services/advisory/ey-etude-big-data-2014-10-freins-identifies>]

1. La collecte de la data encore largement limitée aux canaux traditionnels
2. Les données non structurées, le maillon faible de l'analyse
3. Un manque de compétences analytiques
4. Une carence des outils de traitement des données
5. L'analyse de la data encore (trop) peu orientée vers le prédictif et le temps réel
6. Le manque de transversalité dans la gestion des projets (Big) data
7. L'absence de mesure du ROI des projets (Big) data



# Big Data : les 10 freins au développement (EY)

## □ Selon EY (ex-Ernst&Young), analyse de 2014

[<http://www.ey.com/fr/fr/services/advisory/ey-etude-big-data-2014-10-freins-identifies>]

1. La collecte de la data encore largement limitée aux canaux traditionnels
2. Les données non structurées, le maillon faible de l'analyse
3. Un manque de compétences analytiques
4. Une carence des outils de traitement des données
5. L'analyse de la data encore (trop) peu orientée vers le prédictif et le temps réel
6. Le manque de transversalité dans la gestion des projets (Big) data
7. L'absence de mesure du ROI des projets (Big) data
8. Un manque de sponsorship de la direction générale
9. Un risque majeur pour la fiabilité de la data : la réticence à partager des données personnelles

[D'après une récente étude EY 70 % des consommateurs sont réticents à partager leurs données personnelles avec les entreprises et 49 % affirment qu'ils seront moins enclins à le faire dans les cinq années à venir.]

# Big Data : les 10 freins au développement (EY)

## □ Selon EY (ex-Ernst&Young), analyse de 2014

[<http://www.ey.com/fr/fr/services/advisory/ey-etude-big-data-2014-10-freins-identifies>]

1. La collecte de la data encore largement limitée aux canaux traditionnels
2. Les données non structurées, le maillon faible de l'analyse
3. Un manque de compétences analytiques
4. Une carence des outils de traitement des données
5. L'analyse de la data encore (trop) peu orientée vers le prédictif et le temps réel
6. Le manque de transversalité dans la gestion des projets (Big) data
7. L'absence de mesure du ROI des projets (Big) data
8. Un manque de sponsorship de la direction générale
9. Un risque majeur pour la fiabilité de la data : la réticence à partager des données personnelles
10. Faible sensibilisation aux enjeux de sécurité et de protection de la data  
[30 % estiment ne pas être concernées par les enjeux de protection de la vie privée lors de l'exploitation de leurs données clients]

# Big Data

## Plan

### 1. Introduction

- Prolégomènes
- Les deux catégories de raisonnement
- L'exemple Google Flu
- Histoire
- Les multiples Vs
- La possibilité technologique
- Volumes et production de données
- Quelques projets emblématiques
- Technologie et écosystème
- Les entreprises
- Visualiser les données

"There is a central difference between the old and new economies:  
the old industrial economy was driven by economies of scale;  
the new information economy is driven by the economics of networks..."

*Information Rules, C. Shapiro, Hal R. Varian*

# Le Big Data - incise Information/Connaissance

- **Information** : donnée pertinente pour un objectif
  - Complexe à structurer avec un ordinateur
  - Aspect qualitatif et quantitatif : complexe à analyser
  - Nécessite un consensus pour arriver à un sens
  - Plus difficile à faire circuler
- **Information** : flux de données circulant dans une organisation
  - Une quantité de savoir tacite
  - Une quantité de savoir explicite transformable
- **Information et stratégie** : tout dépend des objectifs
  - Stratégique : bénéficier des media internes/externes pour les atteindre
  - Tactique : s'organiser pour les atteindre
  - Opérationnel : ce qui est essentiel pour les atteindre

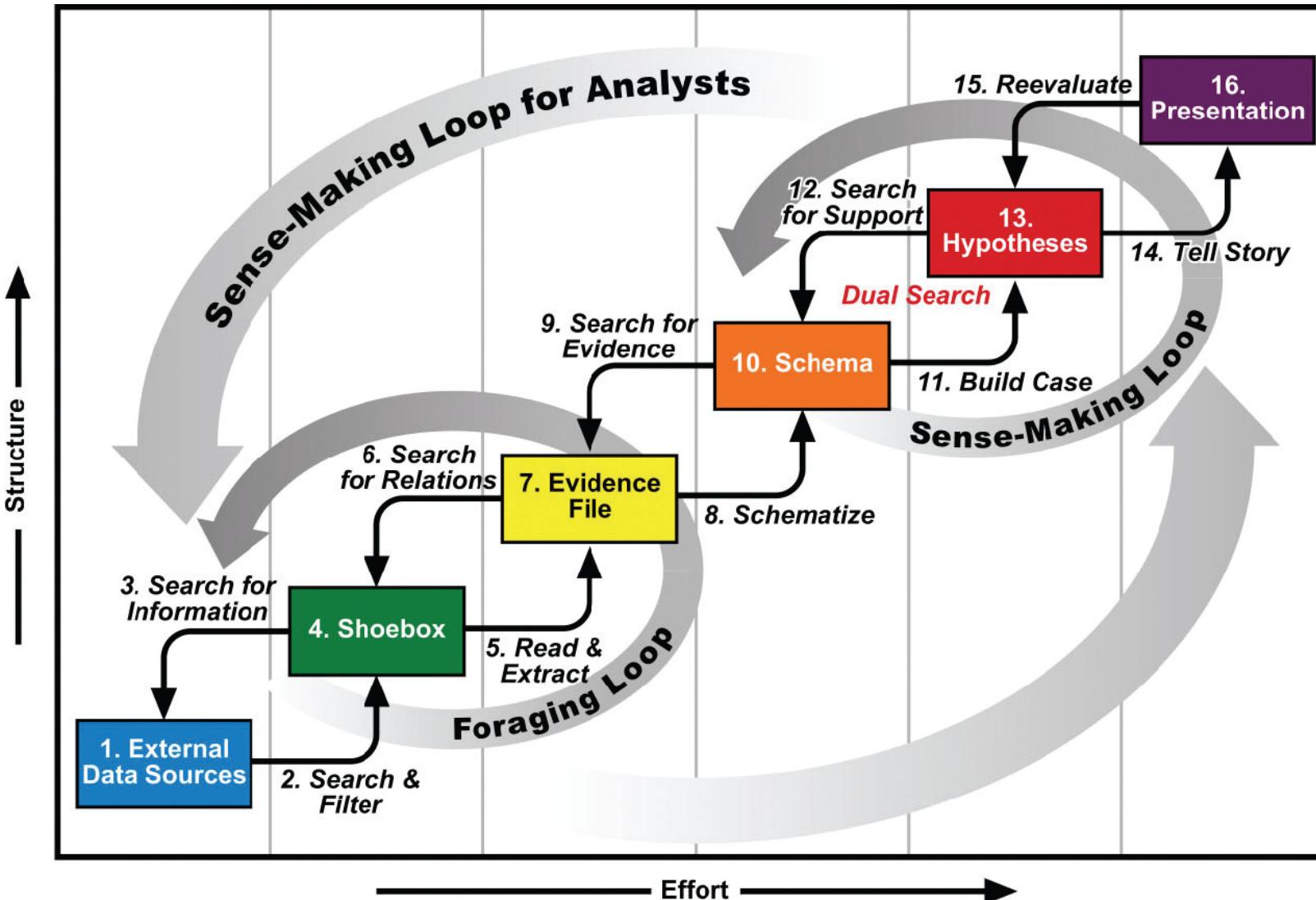
# Le Big Data - incise Information/Connaissance

- **Connaissance** : donnée, information intégrée par le cerveau
  - Très complexe à structurer avec un ordinateur  
(Cf. les sciences cognitives)
  - Interprétation, synthèse et contexte : analytiquement très complexe
  - Encore plus difficile à faire circuler
- **Connaissance**
  - **Tacite** : de l'ordre des représentations mentales
    - Nait et meurt avec les personnes
    - Mémoire (existante et information), contexte (situations et relations), compréhension (association et décision), éthique (croyances et valeurs)
  - **Explicite** ou structurée
    - Mémoire mécanique (documentation, banque de données, livre), contexte structuré (processus organisationnel), connaissance tacite qui peut être stockée et manipulée automatiquement
    - Survit à la personne

# Problématique de l'analyse visuelle (« visual analytics »)

→ Création de sens par...

...la CIA

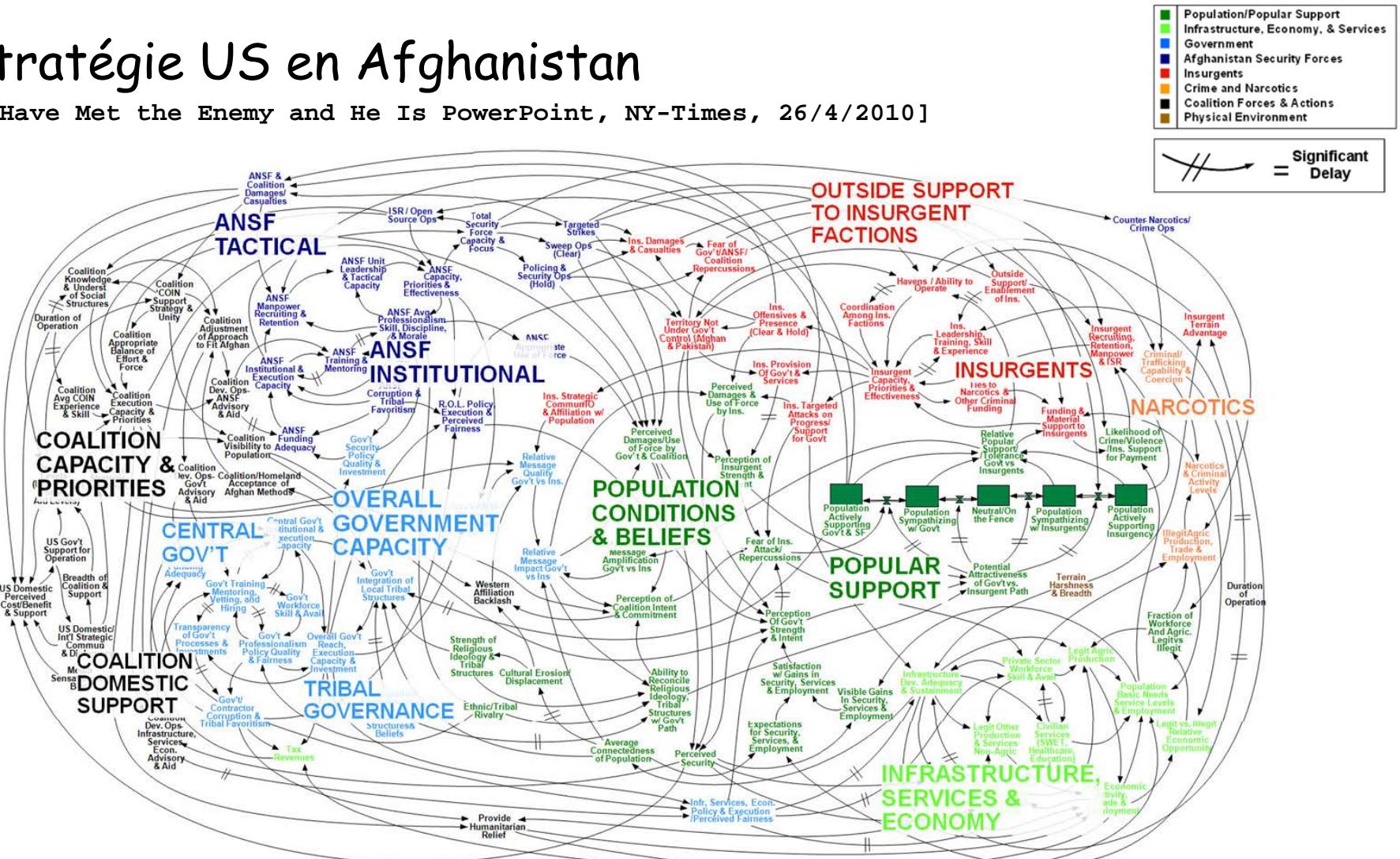


Thomas, J. J. and K. A. Cook, Eds. (2006). *Illuminating the Path: The Research and Development Agenda for Visual Analytics* IEEE Computer Society.

# Importance de la visualisation des données

## Stratégie US en Afghanistan

[We Have Met the Enemy and He Is PowerPoint, NY-Times, 26/4/2010]

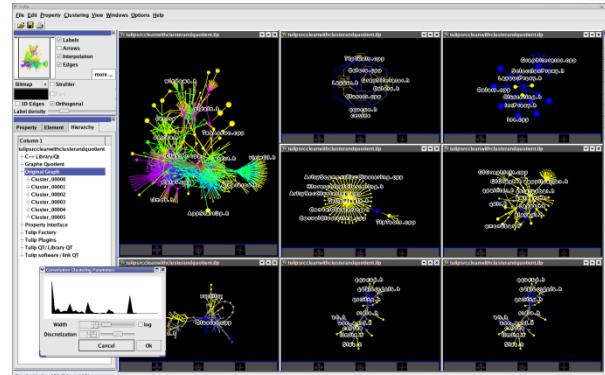


# Valeur ajoutée de la fouille visuelle et interactive

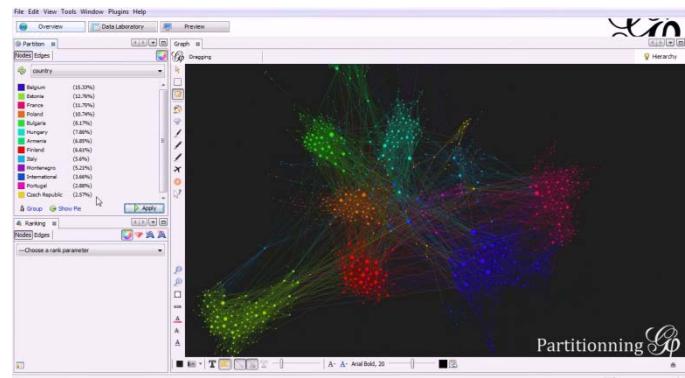
- KDD Panel « The perfect Data Mining Tool » (Ankerst'02)
  - The human eye is an excellent tool for spotting natural patterns
  - Getting rid of the human in the loop? Wrong decision!
  - Increase human participation through visualization in the data exploration and knowledge discovery processes

# Fouille visuelle interactive

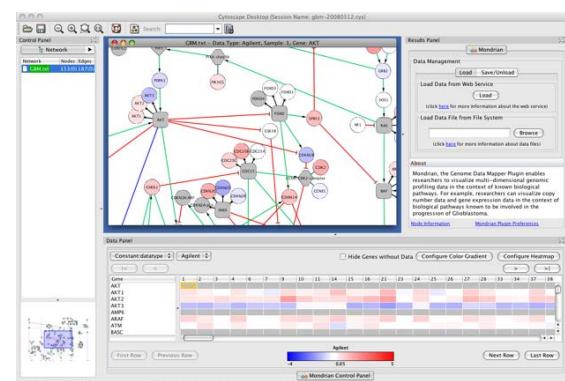
- Tulip  
<http://tulip-software.org/>



- Gephi  
<http://gephi.org>



- Cytoscape  
<http://www.cytoscape.org/>



# Interaction: la vraie valeur ajoutée de la visualisation

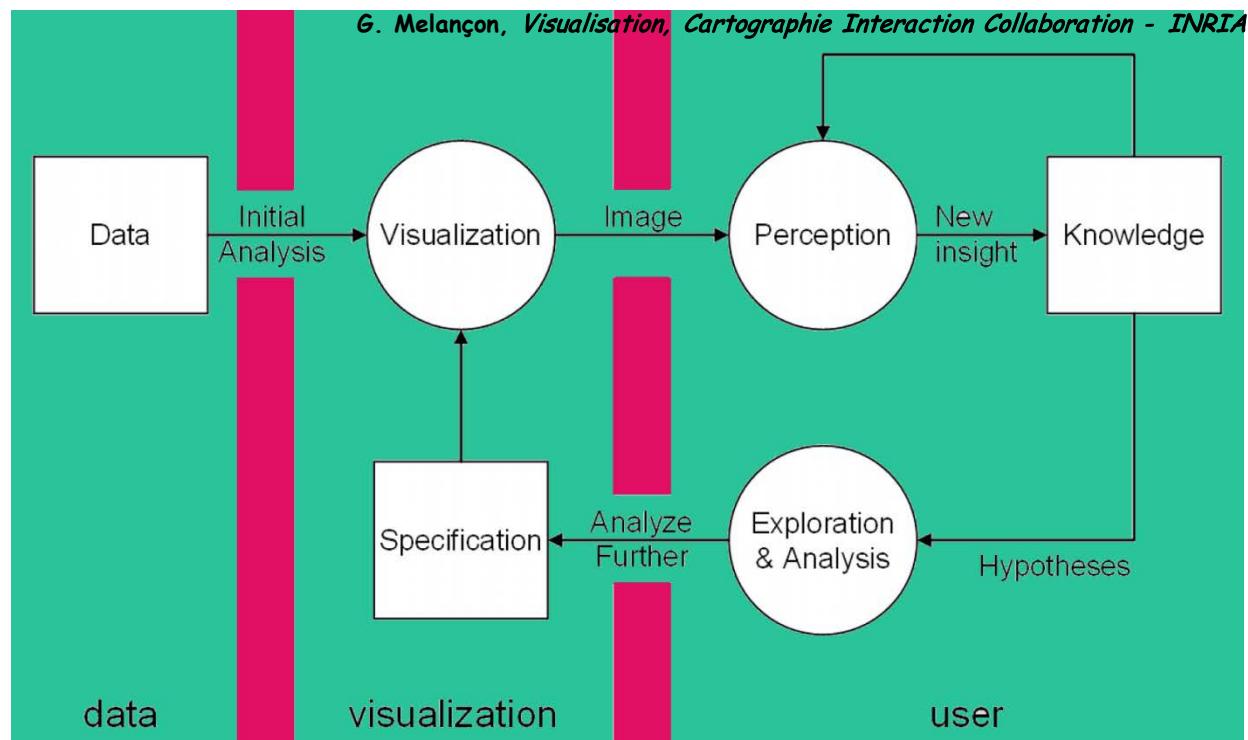
- Questions que ne résoudront pas les avancées de la technologie matérielle
  - Penser la visualisation en « interaction »
  - Bien comprendre quand et pourquoi la visualisation porte ses fruits
  - Collaboration avec les autres domaines d'application
  - L'intégration de la visualisation avec d'autres méthodologies

NIH-NSF Visualization Research Challenges Report, 2006

# Analyse visuelle

## □ [Keim 2006]

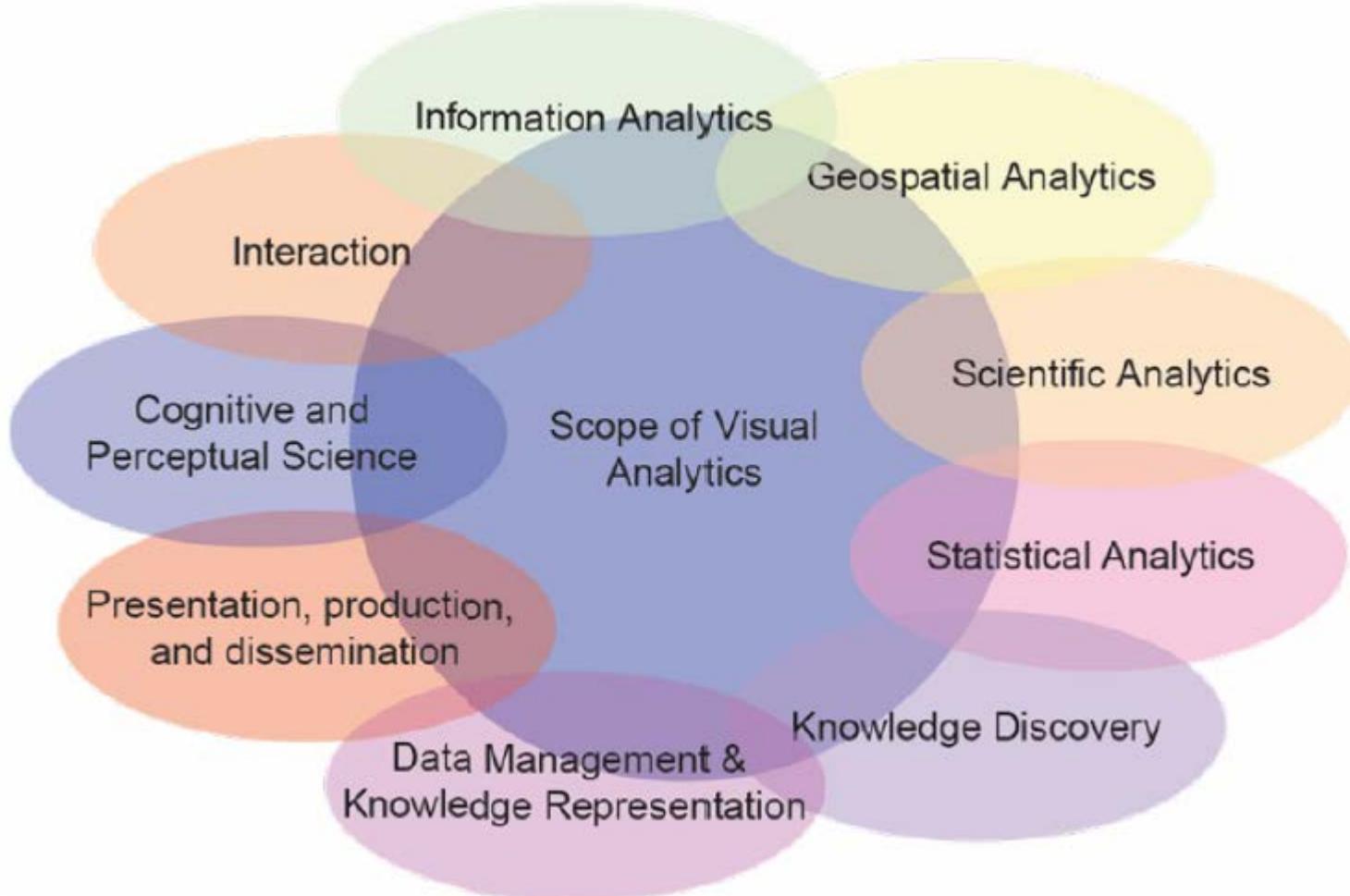
1. Analyze first
2. Show the important
3. Zoom, filter, and analyze furter
4. Details on demand



# Valeur ajoutée de la fouille visuelle

- Vient en appui du travail d'analyse
  - ➔ Exploration
  - ➔ Modélisation
    - En apportant des représentations visuelles et interactives des données
    - Tire parti du système visuel et de la cognition humaine
  - ➔ Analyse exploratoire vs analyse confirmatoire

# Visualisation : un domaine pluri-disciplinaire



Thomas, J. J. and K. A. Cook, Eds. (2006). Illuminating the Path: The Research and Development Agenda for Visual Analytics IEEE Computer Society.

# Interaction et hiérarchisation de l'espace

## Propositions

- Simplification de l'espace fouillé
- Production de vues abstraites
- Nécessaire interaction pour dévoiler les détails

## [Keim 2006]

1. Analyze first
2. Show the important
3. Zoom, filter, and analyze further
4. Details on demand

# Interaction et hiérarchisation de l'espace

## □ Modèles hiérarchiques en sciences

- Réseaux sociaux
- Structure du langage
- Graphes du web
- Réseaux structurels
- Réseaux fonctionnels
- Entités biologiques
- Ensembles logiciels
- ...

# Table périodique des outils de visualisation

→ [http://www.visual-literacy.org/periodic\\_table/periodic\\_table.html](http://www.visual-literacy.org/periodic_table/periodic_table.html)

## A PERIODIC TABLE OF VISUALIZATION METHODS

> < <b>C</b> continuum	<b>Data Visualization</b> Visual representations of quantitative data in schematic form (either with or without axes)												<b>Strategy Visualization</b> The systematic use of complementary visual representations in the analysis, development, formulation, communication, and implementation of strategies in organizations.												
> < <b>Tb</b> table	> < <b>Ca</b> cartesian coordinates	> < <b>Pi</b> pie chart	> < <b>L</b> line chart	> < <b>B</b> bar chart	> < <b>Ac</b> area chart	> < <b>R</b> radar chart cobweb	> < <b>Pa</b> parallel coordinates	> < <b>Hy</b> hyperbolic tree	> < <b>Cy</b> cycle diagram	> < <b>T</b> timeline	> < <b>Ve</b> venn diagram	> < <b>Mi</b> mindmap	< > <b>Sq</b> square of oppositions	> < <b>Cc</b> concentric circles	> < <b>Ar</b> argument slide	> < <b>Co</b> communication diagram	> < <b>Fp</b> flight plan	> < <b>Mm</b> metro map	> < <b>Tm</b> temple	< > <b>St</b> story template	> < <b>Tr</b> tree	> < <b>Ct</b> cartoon			
> < <b>Hi</b> histogram	> < <b>Sc</b> scatterplot	> < <b>Sa</b> sankey diagram	> < <b>In</b> information lense	> < <b>E</b> entity relationship diagram	> < <b>Pt</b> petri net	> < <b>Fl</b> flow chart	< > <b>Cl</b> clustering	> < <b>Lc</b> layer chart	> < <b>Py</b> minto pyramid technique	> < <b>Ce</b> cause-effect chains	> < <b>Tl</b> toulmin map	> < <b>Sw</b> swim lane diagram	> < <b>Gc</b> gantt chart	< > <b>Pm</b> perspectives diagram	> < <b>D</b> dilemma diagram	< > <b>Pr</b> parameter ruler	> < <b>Kn</b> knowledge map	> < <b>Ic</b> iceberg	> < <b>Lm</b> learning map						
> < <b>Tk</b> tukey box plot	> < <b>Sp</b> spectrogram	> < <b>Da</b> data map	> < <b>Tp</b> treemap	> < <b>Cn</b> cone tree	> < <b>Sy</b> system dyn./ simulation	> < <b>Df</b> data flow diagram	< > <b>Se</b> semantic network	> < <b>So</b> soft system modeling	< > <b>Sn</b> synergy map	< > <b>Fo</b> force field diagram	> < <b>Id</b> ibis argumentation map	> < <b>Pr</b> process event chains	< > <b>Ev</b> evocative knowledge map	> < <b>V</b> Vee diagram	< > <b>Hh</b> heaven 'n' hell chart	> < <b>I</b> infomural									
<b>Cy</b> <b>Process Visualization</b>					Note: Depending on your location and connection speed it can take some time to load a pop-up picture. © Ralph Lengler & Martin J. Eppler, <a href="http://www.visual-literacy.org">www.visual-literacy.org</a>																				version 1.5
<b>Hy</b> <b>Structure Visualization</b>					> < <b>Su</b> supply demand curve	> < <b>Pc</b> performance charting	> < <b>St</b> strategy map	> < <b>Oc</b> organisation chart	< > <b>Ho</b> house of quality	> < <b>Fd</b> feedback diagram	> < <b>Ft</b> failure tree	> < <b>Mq</b> magic quadrant	> < <b>Ld</b> life-cycle diagram	> < <b>Po</b> porter's five forces	< > <b>S</b> s-cycle	> < <b>Sm</b> stakeholder map	< > <b>Is</b> ishikawa diagram	> < <b>Tc</b> technology roadmap							
<b>Hy</b> <b>Overview Detail</b>					> < <b>Ed</b> edgeworth box	> < <b>Pf</b> portfolio diagram	> < <b>Sg</b> strategic game board	> < <b>Mz</b> mintzberg's organigraph	< > <b>Z</b> zwicky's morphological box	< > <b>Ad</b> affinity diagram	< > <b>De</b> decision discovery diagram	> < <b>Bm</b> bg matrix	> < <b>Stc</b> strategy canvas	> < <b>Vc</b> value chain	< > <b>Hy</b> hype-cycle	> < <b>Sr</b> stakeholder rating map	> < <b>Ta</b> taps	> < <b>Sd</b> spray diagram							

# Table périodique des outils de visualisation

□ [http://www.visual-literacy.org/pages/maps/mapping\\_tools\\_radar/radar.html](http://www.visual-literacy.org/pages/maps/mapping_tools_radar/radar.html)

