Big Data

Survivre au déluge des données

Aurélien Garivier et Jean-Michel Loubes

Institut de Mathématiques de Toulouse Université Paul Sabatier, Toulouse 3

ML Team Creation Seminar, 2016

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Plan de l'exposé

Les chiffres et les statistiques

Le phénomène Big Data
- Les Big Data dans la presse
- Qu'est-ce que les Big Data

Quels défis scientifiques ?
- Éviter les fausses découvertes
- Tirer partie des opportunités

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Un monde numérique

Nous sommes rentrés dans l'ère du numérique

- Les données sont partout : sondages, indices de popularité, scores électoraux, statistiques économiques voire sportives ....
- Nous produisons nous-mêmes des données de notre plein grès : photos, interactions sur les réseaux sociaux, objets connectés ...
- sans le vouloir vraiment : recherches sur Internet, traces laissées par nos actions (achats, réseaux sociaux ...)
- ou en échange de services (pas si) gratuits.

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Un monde numérique

Notre vie est de plus en plus dirigée par ce monde numérique

- Société de consommation : Algorithmes prédictifs pour prévoir nos achats
- Notre relation avec les banques, assurances : scoring pour déterminer crédit, avantages
- Algorithmes pour la carte scolaire, pour l'orientation scolaire...
- Surveillance policière ...
- Choix d'un conjoint par sites de rencontre ...
- Voitures qui roulent seule et réagissent à leur environnement ...

Tout cela grâce aux mathématiques ... (pas que)

[Src: Loubes, *Petit guide de self-défense contre la statistique et ceux qui la manipulent*]

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

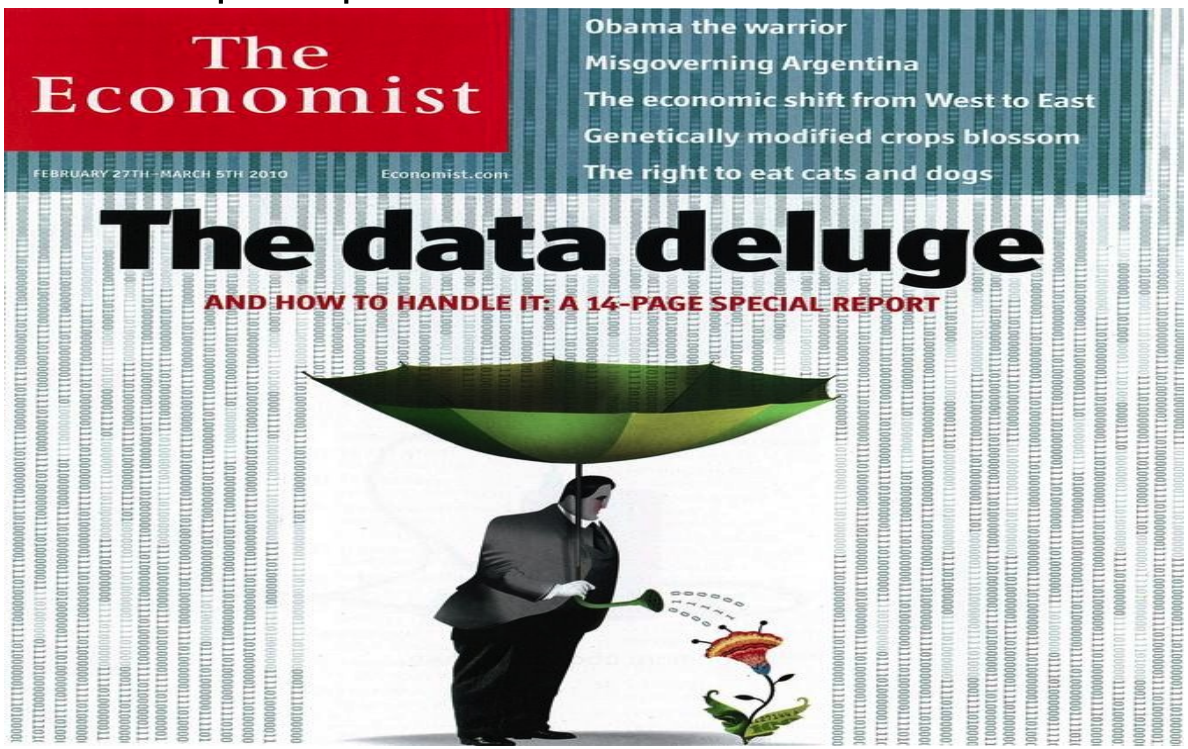
Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Un monde numérique : le mythe du **Big Data**

Des chiffres tous les jours dans tous les media , chiffres sur lesquels s'appuie la pensée :
Economie quantitative, psychologie quantitative mais aussi décisions politiques.



Un monde numérique : points positifs



- Une donnée quantitative (un nombre) ça ne se discute pas

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Un monde numérique : points positifs



- Une donnée quantitative (un nombre) ça ne se discute pas
- c'est sérieux

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Un monde numérique : points positifs



- Une donnée quantitative (un nombre) ça ne se discute pas
- c'est sérieux
- c'est facile à interpréter

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Un monde numérique : points positifs



Bref ... **c'est efficace** !

- Une donnée quantitative (un nombre) ça ne se discute pas
- c'est sérieux
- c'est facile à interpréter
- c'est synthétique ..

Points négatifs

Peut-on faire mentir les nombres ?

- Disraeli : Il y a les mensonges, les gros mensonges, et les statistiques.

Par manque de temps , de connaissances, on s'expose à des erreurs d'interprétation.
Nécessité de comprendre l'usage de la statistique et leur signification : la vérité mathématique est-elle vérité au sens usuel ?

- HG. Wells : Le jugement statistique sera un jour aussi nécessaire à l'exercice de base des fonctions du citoyen que la capacité de lire et d'écrire.

Les chiffres et les statistiques
Le phénomène Big Data
Les Big Data dans la presse
Qu'est-ce que les Big Data
Quels défis scientifiques ?
Éviter les fausses découvertes
Tirer partie des opportunités

La science des données : la statistique

- La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.
  Source : Wikipedia
- La statistique est utilisée dans presque tous les domaines de l'activité humaine : sciences sociales, économie, médecine, biologie, industrie ...

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Toujours plus de données

- Systèmes de stockage (Cloud) et de partage (Open Data, protocoles peer)
- Monde connecté où tout est numérique : Internet, IOT (**Internet of Things** objects connectés)
- Changement de mentalité : créer de la donnée et la faire circuler de la part
  des utilisateurs
  des entreprises.
- Apparition des Méga-entreprises du numériques : Google, Facebook, Apple et toute une économie du numérique (Uber, Netflix, Amazon ...)

Plan de l'exposé

Les chiffres et les statistiques

Le phénomène Big Data
- Les Big Data dans la presse
- Qu'est-ce que les Big Data

Quels défis scientifiques ?
- Éviter les fausses découvertes
- Tirer partie des opportunités

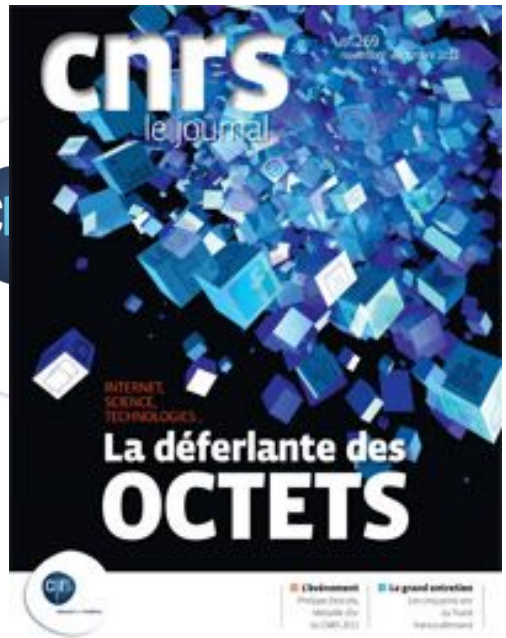
Les chiffres et les statistiques

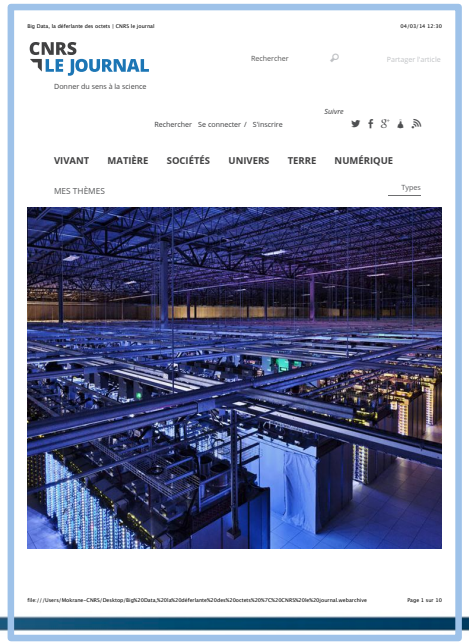
Le phénomène Big Data
- Les Big Data dans la presse
- Qu'est-ce que les Big Data

Quels défis scientifiques ?
- Éviter les fausses découvertes
- Tirer partie des opportunités

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Sur le site du CNRS







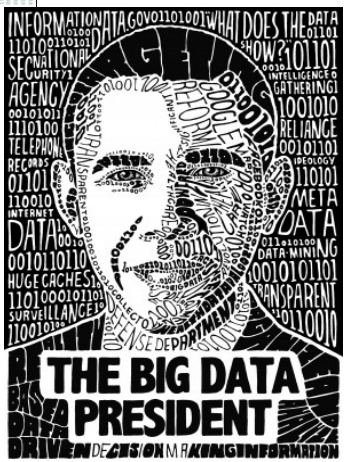
[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Et ailleurs









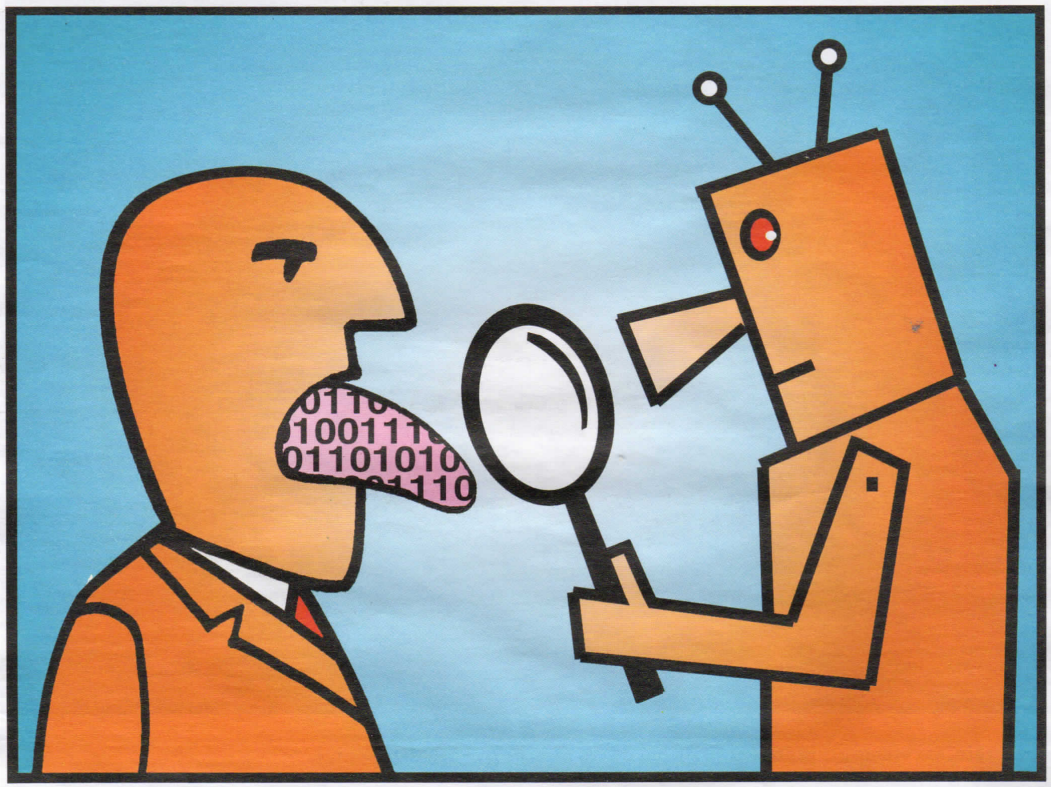
[Src: Bouzeghoub, *L'exploitation scientifique des données*]

40 ◆ Courrier international | n° 1130 | du 28 juin au 4 juillet 2012

En couverture

Nouvelles technologies

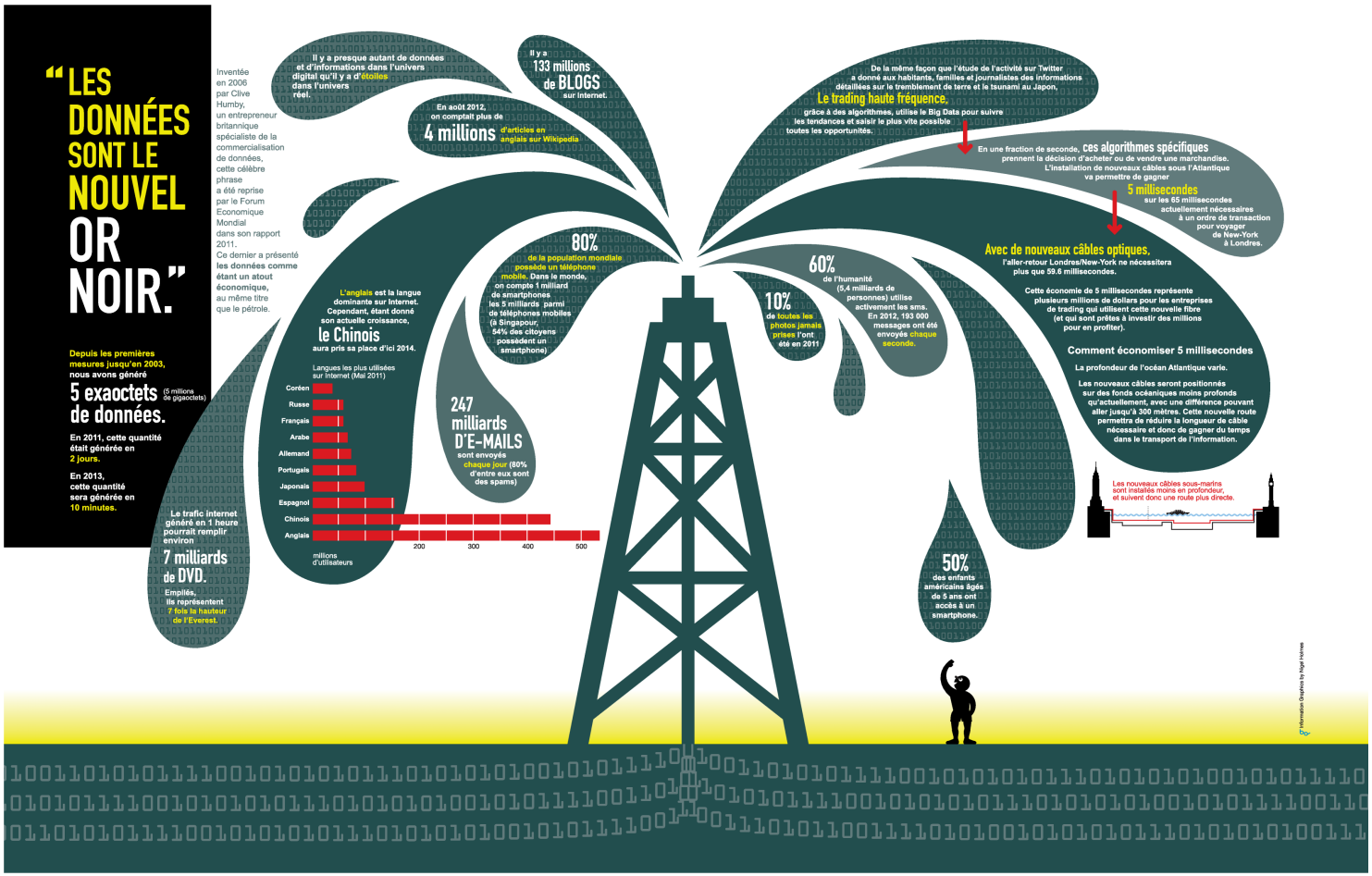
Le partage des données nous sauvera-t-il ?



↑ *Dessin de* **Kopelnitsky,** *Etats-Unis.*

[Src: Loubes, *Petit guide de self-défense contre la statistique et ceux qui la manipulent*]

Croissance du monde numérique



Les chiffres et les statistiques
Le phénomène Big Data
Les Big Data dans la presse
Qu'est-ce que les Big Data
Quels défis scientifiques ?
Éviter les fausses découvertes
Tirer partie des opportunités

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Aucun domaine n'échappe à l'avalanche des données



- **Commerce et les affaires**
  - SI d'entreprise, Banques, transactions commerciales, systèmes de réservation**, ...**
- **Gouvernements et organisations**
  - Lois, réglementations, standards, infrastructures, ….
- **Loisirs**
  - Musique, vidéo, jeux, réseaux sociaux…
- **Sciences fondamentales**
  - Astronomie, physique et énergie, génome, …
- **Santé**
  - Dossier médical, sécurité sociale,…
- **Environnement**
  - Climat, dév durable, pollution, alimentation,…
- **Humanités et Sciences Sociales**
  - Numérisation du savoir (littérature, histoire,art, srchitectures), données archéologiques…

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Qu'est-ce qu'une (très grande) masse de données ?



VLDB

XLDB

Big Data

Very Big Data

Data Deluge

Massive Data

Data Masses

Data inflation 2

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

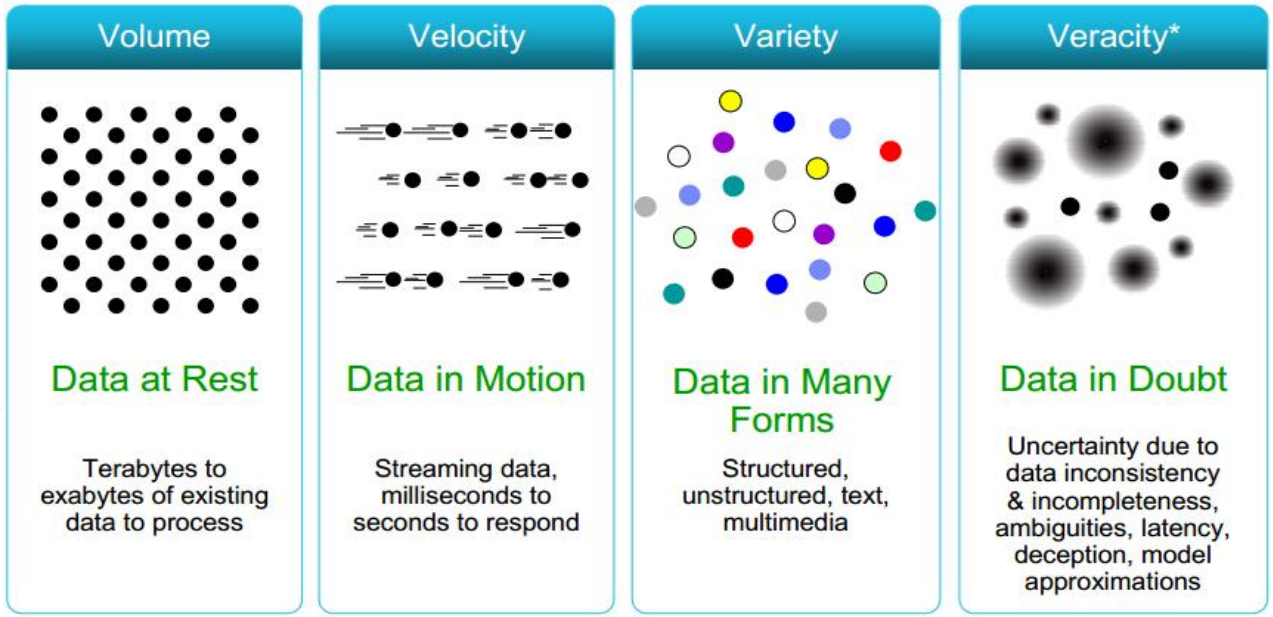
Source: *The Economist*

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

Grandes Conf du domaine: VLDB, XLDB, ICDE, EDBT, …

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Complexité multidimensionnele des Big Data



www.cnrs.fr

- Nouvelles archi. de stockage
- Nouvelles archi. d'interopérabilité

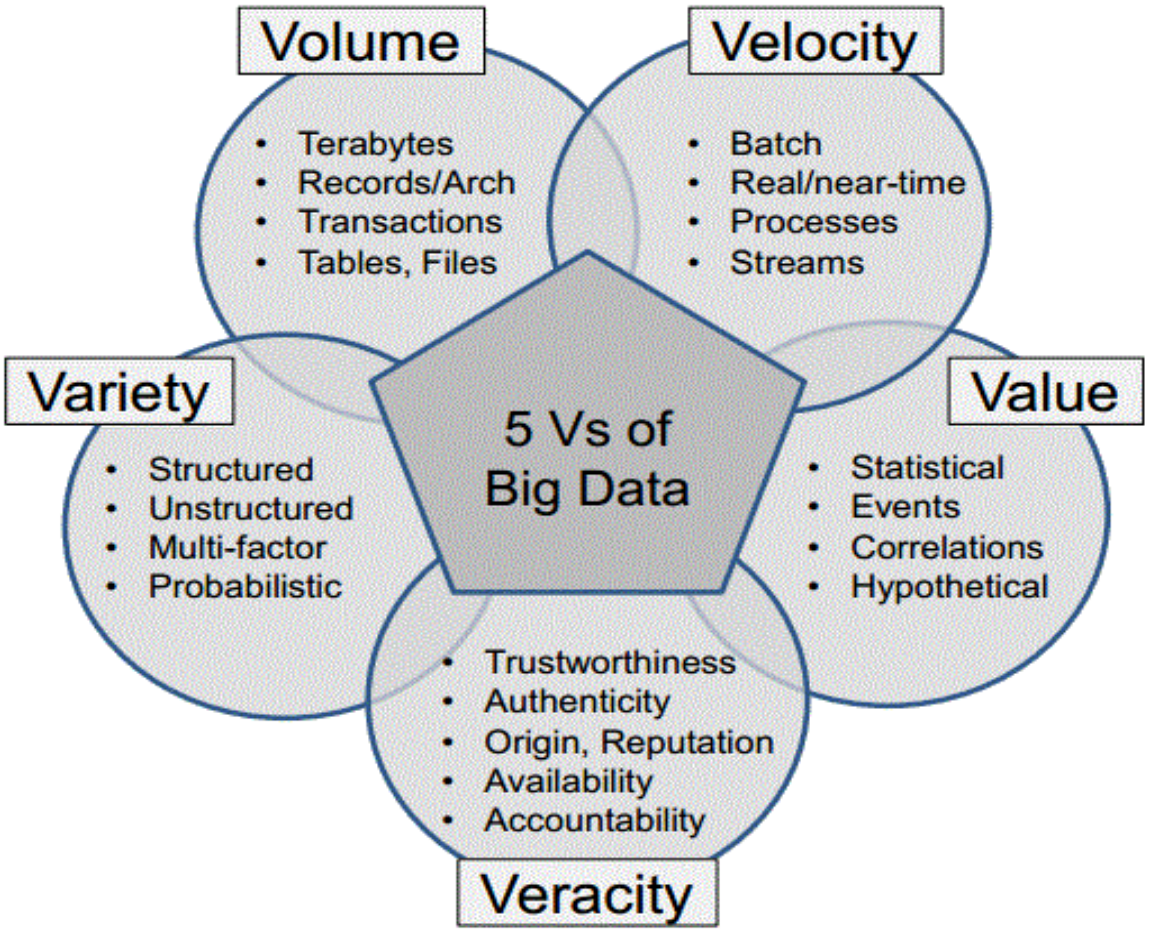
- Défi pour les réseaux de communication
- Nouveaux modèles de calcul sur des flux

- Nettoyage et transformation
- Fusion de données

Nouveaux modèles de qualité (données & processus de traitement)

http://www.datasciencecentral.com/profiles/blogs/data-veracity

Caractéristiques des données : les 3V ou les 5 V



Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

INS2I

Institut des sciences informatiques et de leurs interactions

Une petite idée des volumes de données





Volume	Exemple
1 gigabyte 10^9	Information known in the human genome
1 terabyte 10^{12}	Annual world literature production
1 petabyte 10^{15}	All US academic research libraries
1 exabyte 10^{18}	Two thirds of annual world production of information

Source: http://www.jisc.ac.uk/publications/briefingpapers/2004/pub_datadeluge.aspx

Plan de l'exposé

Les chiffres et les statistiques

Le phénomène Big Data

- Les Big Data dans la presse
- Qu'est-ce que les Big Data

Quels défis scientifiques ?

- Éviter les fausses découvertes
- Tirer partie des opportunités

Les chiffres et les statistiques

Le phénomène Big Data

- Les Big Data dans la presse
- Qu'est-ce que les Big Data

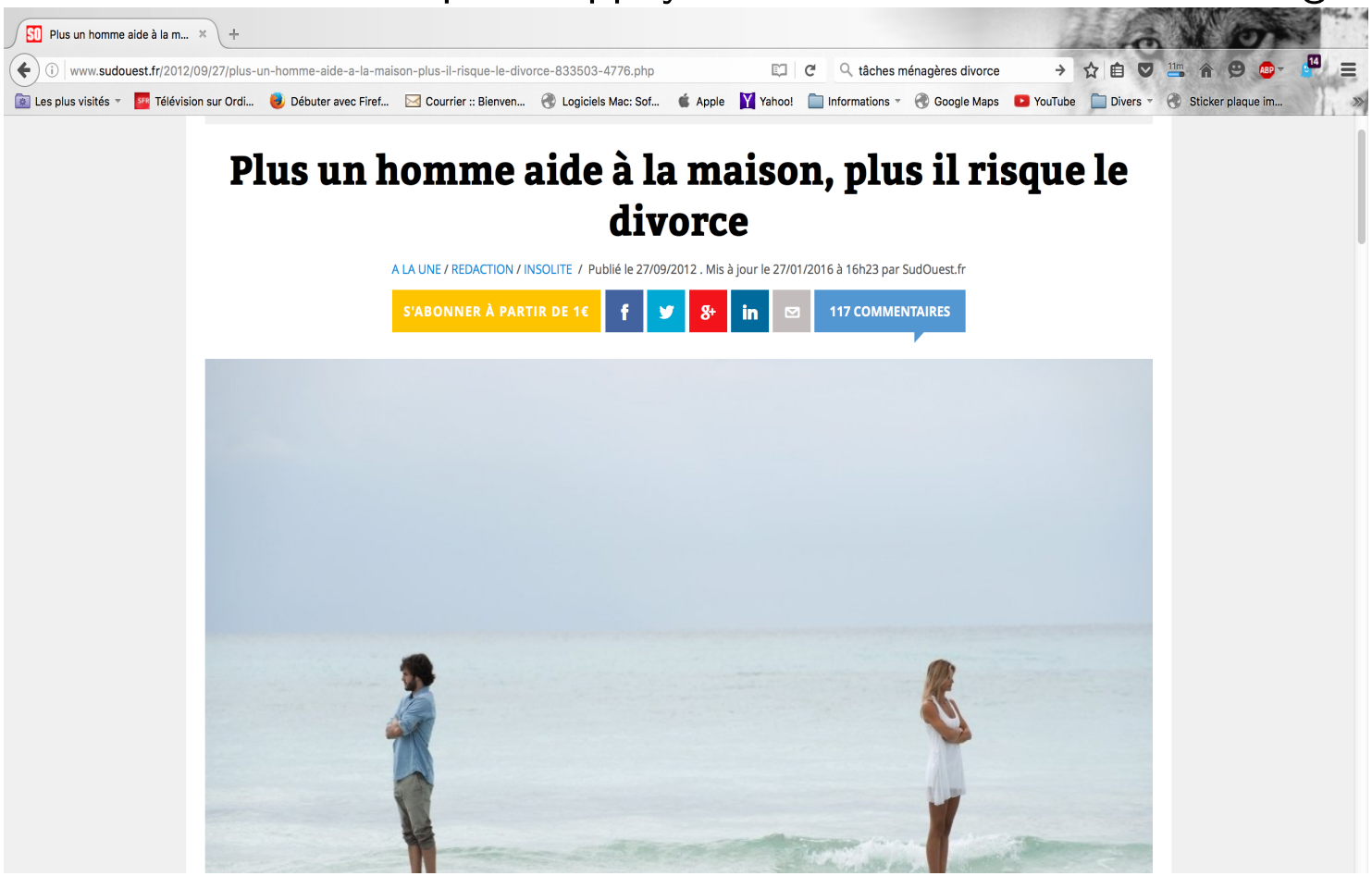
Quels défis scientifiques ?

- Éviter les fausses découvertes
- Tirer partie des opportunités

[Src: Loubes, *Petit guide de self-défense contre la statistique et ceux qui la manipulent*]

Corrélations en statistique

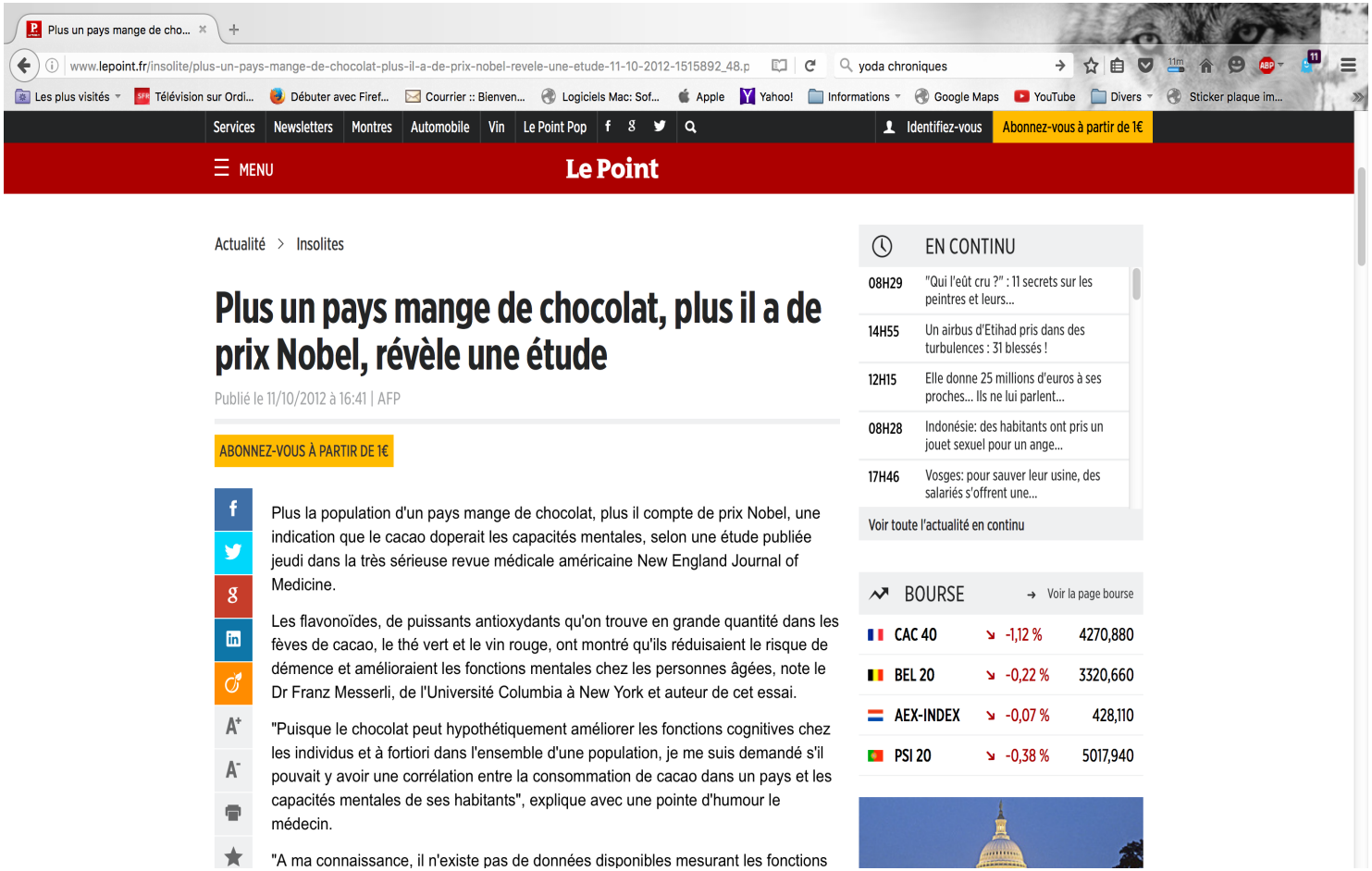
De nombreux journaux parlent d'études scientifiques s'appuyant sur des études et des sondages



[Src: Loubes, *Petit guide de self-défense contre la statistique et ceux qui la manipulent*]

Corrélations en statistique

De nombreux journaux parlent d'études scientifiques s'appuyant sur des études et des sondages



Plus un pays mange de cho...

www.lepoint.fr/insolite/plus-un-pays-mange-de-chocolat-plus-il-a-de-prix-nobel-revele-une-etude-11-10-2012-1515892_48.p

yoda chroniques

Les plus visités | Télévision sur Ordi... | Débuter avec Firef... | Courrier :: Bienven... | Logiciels Mac: Sof... | Apple | Yahoo! | Informations | Google Maps | YouTube | Divers | Sticker plaque im...

Services | Newsletters | Montres | Automobile | Vin | Le Point Pop | Identifiez-vous | Abonnez-vous à partir de 1€

MENU

Le Point

Actualité > Insolites

Plus un pays mange de chocolat, plus il a de prix Nobel, révèle une étude

Publié le 11/10/2012 à 16:41 | AFP

ABONNEZ-VOUS À PARTIR DE 1€

Plus la population d'un pays mange de chocolat, plus il compte de prix Nobel, une indication que le cacao doperait les capacités mentales, selon une étude publiée jeudi dans la très sérieuse revue médicale américaine New England Journal of Medicine.

Les flavonoïdes, de puissants antioxydants qu'on trouve en grande quantité dans les fèves de cacao, le thé vert et le vin rouge, ont montré qu'ils réduisaient le risque de démence et amélioraient les fonctions mentales chez les personnes âgées, note le Dr Franz Messerli, de l'Université Columbia à New York et auteur de cet essai.

"Puisque le chocolat peut hypothétiquement améliorer les fonctions cognitives chez les individus et à fortiori dans l'ensemble d'une population, je me suis demandé s'il pouvait y avoir une corrélation entre la consommation de cacao dans un pays et les capacités mentales de ses habitants", explique avec une pointe d'humour le médecin.

"A ma connaissance, il n'existe pas de données disponibles mesurant les fonctions

EN CONTINU

08H29	"Qui l'eût cru ?" : 11 secrets sur les peintres et leurs...
14H55	Un airbus d'Etihad pris dans des turbulences : 31 blessés !
12H15	Elle donne 25 millions d'euros à ses proches... Ils ne lui parlent...
08H28	Indonésie: des habitants ont pris un jouet sexuel pour un ange...
17H46	Vosges: pour sauver leur usine, des salariés s'offrent une...

Voir toute l'actualité en continu

BOURSE → Voir la page bourse

CAC 40	-1,12 %	4270,880
BEL 20	-0,22 %	3320,660
AEX-INDEX	-0,07 %	428,110
PSI 20	-0,38 %	5017,940

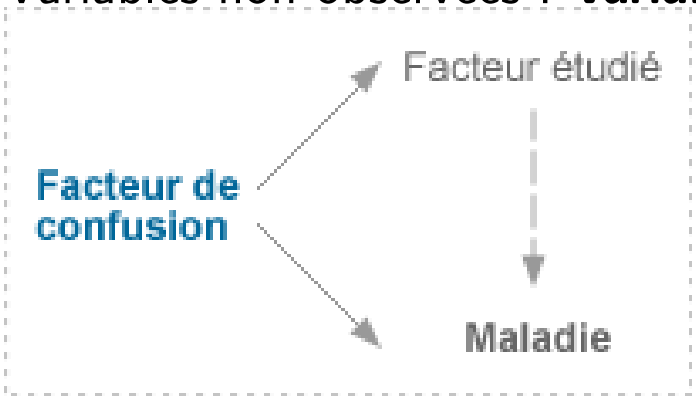


Incompréhension autour de la notion de corrélation

- En grandes dimensions **tout est corrélé**.
- **Corrélation n'est pas causalité** :
  A nombre de prix Nobel et B consommation de chocolat

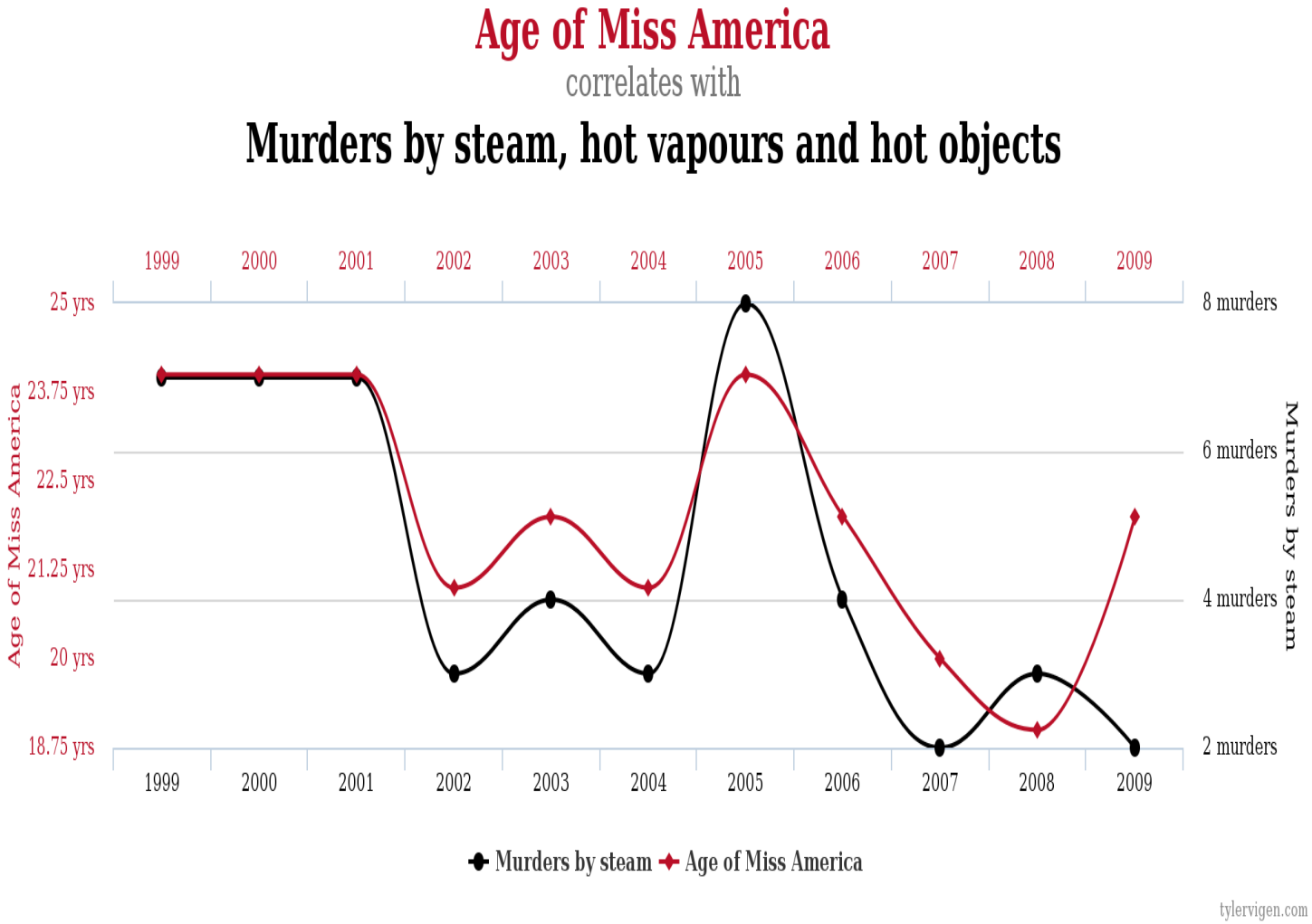
$$A \mapsto B \quad \text{ou} \quad B \mapsto A$$

- Variables non observées : **variables confondantes**



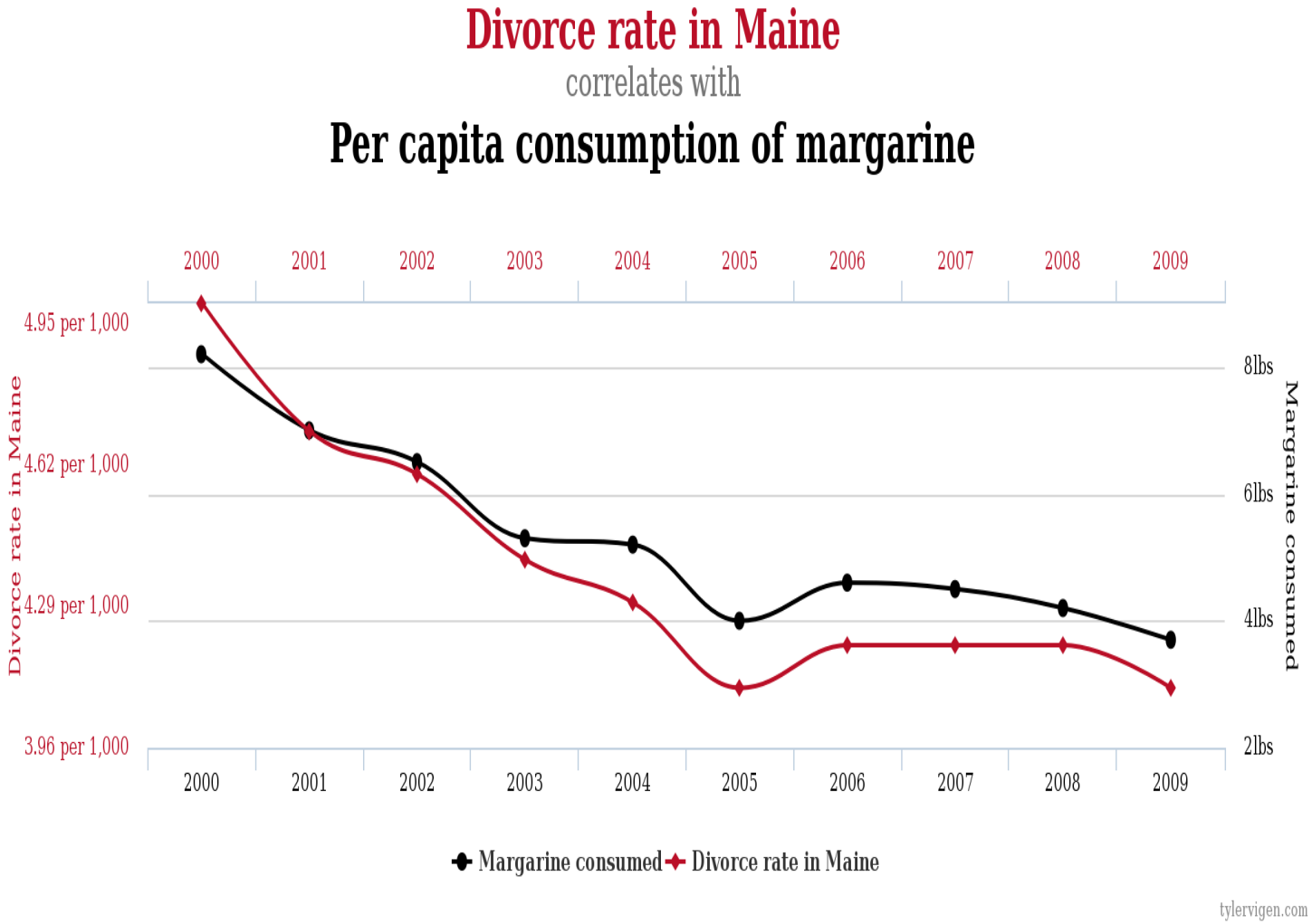
Les chiffres et les statistiques
Le phénomène Big Data
Les Big Data dans la presse
Qu'est-ce que les Big Data
Quels défis scientifiques ?
Éviter les fausses découvertes
Tirer partie des opportunités

[Src: Loubes, *Petit guide de self-défense contre la statistique et ceux qui la manipulent*]



Les chiffres et les statistiques
Le phénomène Big Data
Les Big Data dans la presse
Qu'est-ce que les Big Data
Quels défis scientifiques ?
Éviter les fausses découvertes
Tirer partie des opportunités

[Src: Loubes, *Petit guide de self-défense contre la statistique et ceux qui la manipulent*]



Les chiffres et les statistiques
Le phénomène Big Data
Les Big Data dans la presse
Qu'est-ce que les Big Data
Quels défis scientifiques ?
Éviter les fausses découvertes
Tirer partie des opportunités

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Les grands challenges scientifiques du Big Data

Stockage et préservation des données

- Performance des accès, disponibilité des données
- Protection des données
- Indexation sémantique (ontologies), indexation participative (folksonomies)

Analyse statistique et sémantique, raisonnement

- Analyse en temps réel de flux continus de données émanant de différentes sources
- Requêtes multidimensionnelles sur des grands ensembles de données
- Extraction et interprétation de connaissances

Impact sociétal et économique

- Protection de la vie privée, Droit à l'oubli
- Droits de propriétés, droits d'exploitation
- **Economie d'énergie**
  - **coût du stockage, coût de transfert**

→ 120 kWh/an/TO stocké par CCIN2P3
→ 1M€ /an facture électricité de l'IDRIS

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Exemple 1 : Analyse complexe à grande échelle

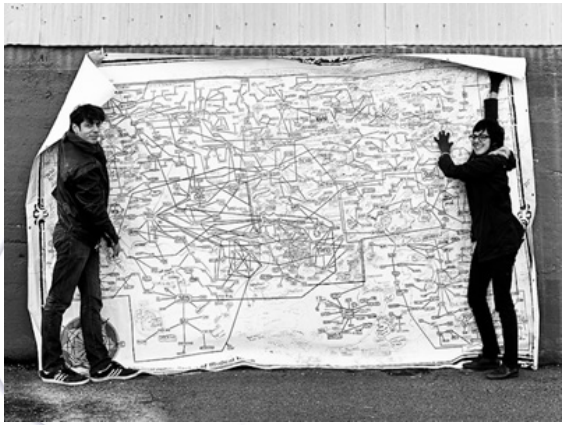
- **Analyse en temps réel de flots continus de données émanant de différentes sources**
  - **Ex: Découvrir et comprendre les patterns caractéristiques du comportement de certains phénomènes ou certaines populations**
- **Réaction en temps réel à des événements d'alerte**
  - **Ex: attaques sur le réseau**
- **Requêtes multidimensionnelles sur des grands ensembles de données**
  - **Découvrir des corrélations entre phénomène**

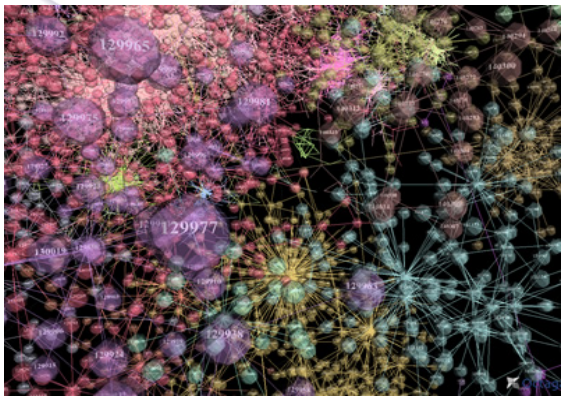
www.cnrs.fr



[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Exemple 2 : La visualisation des données





visualcomplexity.com/vc

- **Besoins**
  - Navigation intuitive/contextuelle
  - Visualisation de phénomènes non perceptibles (durant la simulation)
  - Analyse /Interaction visuelle
- **Problèmes**
  - L'approche de visualisation peut-elle aider à la compréhension d'un phénomène
  - Ou peut-elle introduire un biais et en altérer l'interprétation?
- **Ex: Visualisation post-traitement V.S Visualisation in-situ**
  - Évite des zones d'ombre par perte de calculs intermédiaires
  - → Coupler la simulation et la visualisation (vars température, pression,...)
  - Ex: Swiss Nat Supercomputing Center

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Exemple 3 : La préservation des données







- **Comment préserver les données à durée de vie illimité?**
  - connaissances scientifiques
  - produits culturelles
  - connaissances archéologiques et environnementales
  - connaissances sociales (recensements)
- **Comment préserver les données à durée de vie longue mais limitée**
  - patrimoine informationnel des entreprises
  - Données personnelles (stockées dans les disques privés ou publiés sur le Web)
  - Données publiques (fichiers sécu, police, ...)
- **Quel coût pour la préservation des données**
  - Coût de conversion des données (formats)
  - Coût pour la migration des technologies
  - Coût de maintien des technologies de niche
- **Quelle stratégie pour les données gérées dans le Cloud ?**

Les chiffres et les statistiques

Le phénomène Big Data

Les Big Data dans la presse

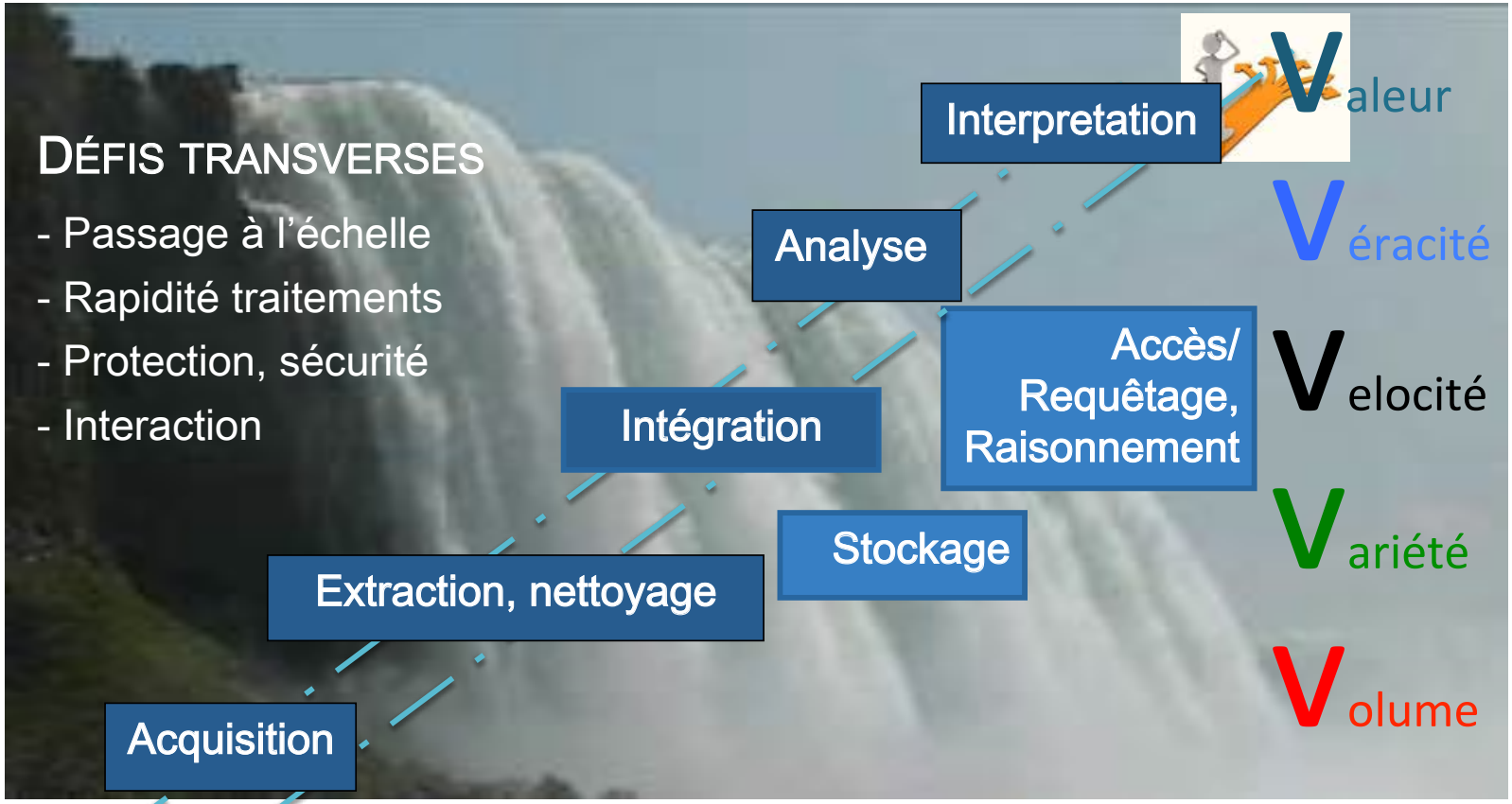
Qu'est-ce que les Big Data

Quels défis scientifiques ?

Éviter les fausses découvertes

Tirer partie des opportunités

Défis accompagnant les chgts



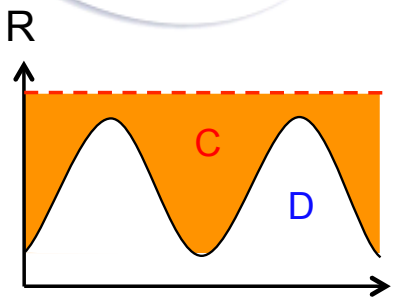
repenser les outils algorithmiques et mathématiques

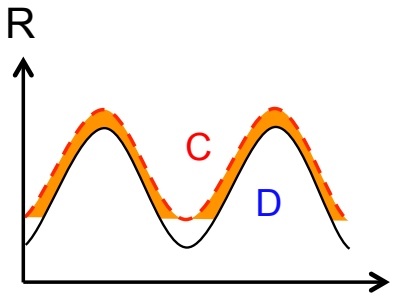
inspired by "Big Data and Its Technical Challenges, Communications of the ACM, July 2014, vol 57, n°7", © H.V. Jagadish et all.

INS2I
Institut des sciences informatiques et de leurs interactions

Défi 1 : Stockage à grande échelle (Cloud)

- **Bénéfices du Cloud**
  - **Pas d'infrastructure à acquérir ni à gérer**
    - ✓ « Pay as you go »
  - **Stockage massif de données (à moindre coût)**
    - ✓ Coûts de stockage et d'utilisation réduits
  - **Accès anytime – anywhere via Internet**
    - ✓ Ex: iCloud (Apple)
  - **Qualité de service**
    - ✓ Disponibilité, sécurité
  - **Elasticité**
    - ✓ absorbe facilement les charges lourdes ou soudaines





- **Challenges du Cloud**
  - **Indexation intelligente (sémantique)**
  - **Sécurité et Confidentialité (privacy)**
  - **Calcul haute performance (//)**
  - **Cohérence et qualité des données**

INS2I
Institut des sciences informatiques et de leurs interactions

Défi 2 : Analyse complexe à grande échelle

- **Analyse en temps réel de flots continus de données émanant de différentes sources**
  - **Ex: Découvrir et comprendre les patterns caractéristiques du comportement des clients/ utilisateurs**
- **Réaction en temps réel à des événements d'alerte**
  - **Ex: attaques sur le réseau**
- **Requêtes multidimensionnelles sur des grands ensembles de données**
  - **Découvrir et comprendre des patterns en analysant le comportement d'une population**
  - **Découvrir des corrélations entre phénomènes**

Divy Agrawal et al
VLDB Tutorial'2010

Big Data
Aurélien Garivier et Jean-Michel Loubes
Les chiffres et les statistiques
Le phénomène Big Data
Les Big Data dans la presse
Qu'est-ce que les Big Data
Quels défis scientifiques ?
Éviter les fausses découvertes
Tirer partie des opportunités

[Src: Bouzeghoub, *L'exploitation scientifique des données*]

INS2I

Institut des sciences informatiques et de leurs interactions

Défi 3 : la gestion de flots d'événements

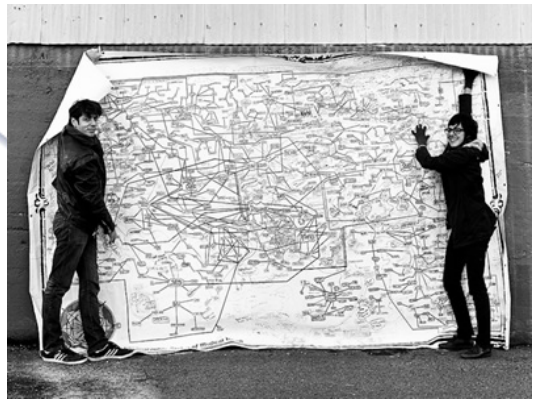
- **Capture d'événements**
  - Politique de détection et de composition (requêtes sur les évnts)
  - Introduction d'incertitude sur l'arrivée des événements et sur le contenu de leurs messages
  - Détection / simulation d'événements rares
- **Réaction aux événements**
  - Politique de déclenchement, consommation, exécution (optimisation)
  - Couplage transactionnel
  - Analyse : confluence, terminaison
- **Bufferisation**
  - Taille des fenêtres temporelles
  - Stratégies de glissement
- **Historisation**
  - Stockage massif
  - Indexation
  - **Analyse complexe** (analyse de séquences, motifs fréquents,…)

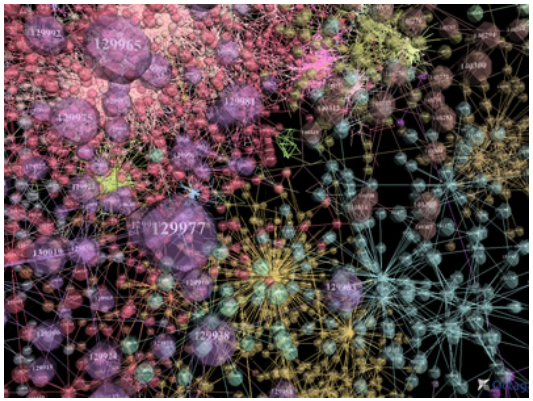
Cible : intelligence ambiante, réseaux sociaux, surveillance temps réel, robotique, bioinformatique.

INS2I

Institut des sciences informatiques et de leurs interactions

Défi 4 : La visualisation des masses de données





- **Besoins**
  - **Navigation intuitive/contextuelle**
    - ✓ desktop intelligent, réseaux sociaux, contenus MM
  - **Visualisation de phénomènes non perceptibles**
    - ✓ Génome, trou noir
  - **Analyse visuelle**
    - ✓ Découverte de connaissances
- **Challenges**
  - **Invention de nouvelles métaphores graphiques**
  - **Algorithmes de graphes performants (//)**
    - ✓ Optimisation de la visulaisation de grands graphes
  - **Clusterisation et stats de graphes**
  - **Adaptation aux terminaux /équipements**

[Src: Bouzeghoub, *L'exploitation scientifique des données*]

INS2I

Institut des sciences informatiques et de leurs interactions

Défi 5 : La préservation des données

- **Comment préserver les données à durée de vie illimité?**
  - connaissances scientifiques
  - produits culturelles
  - connaissances archéologiques et environnementales
  - connaissances sociales (recensements)
- **Comment préserver les données à durée de vie longue mais limitée**
  - patrimoine informationnel des entreprises
  - Données personnelles (stockées dans les disques privés ou publiés sur le Web)
  - Données publiques (fichiers sécu, police, ...)
- **Quel coût pour la préservation des données**
  - Coût de conversion des données (formats)
  - Coût pour la migration des technologies
  - Coût de maintien des technologies de niche
- **Quelle stratégie pour les données gérées dans le Cloud ?**

Accroissement de l'hétérogénéité des MDD

Plus grandes difficultés pour leur intégration et leur exploitation

INS2I
Institut des sciences informatiques et de leurs interactions

Conclusion

- **Fin de l'ère « One Size Fits All »**
  - Il faut offrir des architectures de données flexibles, avec des services de gestion de données adaptables à chaque type d'application/type de données
- **Les SGBD ne sont plus visibles en tant que systèmes intégrés et cohérents**
  - les services de gestion de données sont enfouis dans des systèmes à plus forte valeur ajoutée (services métiers)
- **Grands challenges des MDD**
  - Passage à l'échelle et calcul haute performance
  - Virtualisation et indexation sémantique
  - Requêtage, extraction des connaissances et visualisation
  - Sécurité et protection de la vie privée