

# **Big Data :**

## **Qu'est ce que ça veut dire ?**

**Pascal Poncelet**

LIRMM - UM

[Pascal.Poncelet@lirmm.fr](mailto:Pascal.Poncelet@lirmm.fr)

<http://www.lirmm.fr/~poncelet>



# Un petit exemple

---



The screenshot shows a red header bar with the 'france inter' logo, a 'LE DIRECT' button, a 'RÉÉCOUTER' button, and a 'répondre' (reply) link. Below the header, a comment is displayed:

 **Guirec (anonyme),**  
mardi 29 mars 2016 à 14:48

Je suis très choqué par ce que nous vend Mathieu Roche. Avant de le vendre à la Chine, à la NSA etc...  
L'analyse des sentiments exprimés dans les sms...selon moi la modélisation est une atteinte grave à nos libertés sentimentales. Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

[répondre](#)



# Une petite analyse manuelle

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Quelle est l'opinion exprimée ?
- M. Guirec n'aime pas les travaux de recherche de M. Roche sur l'analyse des sentiments dans les SMS !



# Un traitement informatique

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Comment détecter l'opinion ?
- Question 1 : le texte est-il positif ou négatif ?
- Question 2 : qui est l'auteur des critiques et le sujet de la critique ?
- Question 3 : quel est l'objet de la critique ?



# Opinion du document

---

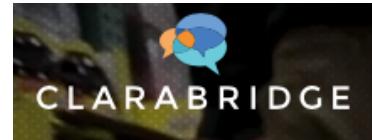
*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Nécessité de faire appel à des ressources externes :
  - Senticnet, Babelnet, Dictionnaires spécialisés, Ontologies



# Opinion du document

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Détection simple des éléments négatifs.  
Même très négatifs (adverbe « Très »).  
Possibilité de prendre en compte une intensité



# Opinion du document

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Vendre est positif ou négatif ... selon le contexte !
- “Libertés sentimentales, poésie” ... positif



# Opinion du document

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Positif, négatif, neutre ?



# Opinion du document

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS selon moi est une atteinte grave à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Pour résumer :
- Quelle fonction d'agrégation pour dire si le document est positif ou négatif ?
- Besoin d'autres informations



# Auteur et sujet de la critique

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

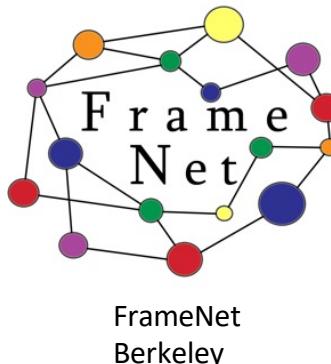
*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Nécessité de faire appel à des ressources externes :
  - Cible du sentiment, source du sentiment



CoreNLP  
Stanford



Treetagger



# Auteur et sujet de la critique

---

*Guirec (Anonyme)* Je suis très choqué par ce que nous vend Mathieu Roche.  
Avant de le vendre à la Chine, à la NSA, etc.

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Nécessité d'avoir une analyse plus fine de la phrase
- *Ce que nous vend ?*
- *L'analyse des sentiments exprimés dans les SMS ...  
selon moi est une atteinte grave à nos libertés  
sentimentales.*



# Conclusion partielle

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Pour l'instant il semble qu'il y ait une opinion négative exprimée par M. Guirec ?
- Une ontologie spécifique pourrait montrer que poésie et sentiments sont proches et montrer que c'est à propos de la poésie (ressource externe spécifique – généralisation de l'approche ?)



# Informations non prises en compte

---

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave  
à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Il est possible de repérer des éléments de géolocalisation ou de détecter des noms propres.
- Nécessité de ressources externes



# Informations non prises en compte

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Mathieu Roche.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.*

*Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Des résultats obtenus de GeoNames

1  Poète

[Republic of the Congo](#), Kouilou

2  Nsa

[Republic of the Congo](#),  
Plateaux

- Des patterns spécifiques de géolocalisation « A la, chez les »



# Attention aux interprétations !

*Guirec (Anonyme) Je suis très choqué par ce que nous vend Michel Roy.*

*Avant de le vendre à la Chine, à la NSA, etc.*

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.*

*Michel Roy est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- M. Guirec exprime une opinion négative !
- Il s'avère que la cible était difficile à définir
- Avec des ressources extérieures (souvent nécessité de prendre en compte des aspects de géolocalisation)



Attention toutefois ...

# Les dérives potentielles

---

- M. Guérec a un avis très négatif sur le Congo (qui est subventionné par la Chine) et il exprime ses sentiments

## **La Chine, premier partenaire du Congo-Brazzaville**

La présence chinoise au Congo, ce n'est pas que des petites boutiques bon marché, loin s'en faut. La Chine est de plus en plus impliquée dans tous les grands travaux de modernisation du pays. D'un coût global de 280 millions de dollars US, la centrale hydraulique d'Imboulou est par exemple le fruit de la coopération sino-congolaise. Idem pour le tout nouvel aérogare et le projet de deuxième piste de

- Nécessité de ressources externes évidentes, nécessité d'autres données, .... nécessité de contrôler



# Et pourtant



# Origine un tweet d'Associated Press

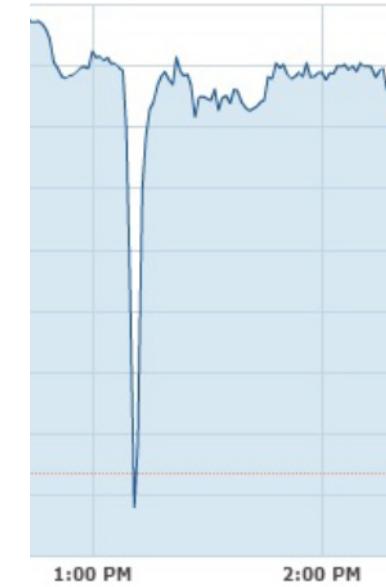
The Associated Press Following  
@AP

Breaking: Two Explosions in the White House and Barack Obama is injured

Reply Retweet Favorite More

3,146 RETWEETS 149 FAVORITES

1:07 PM - 23 Apr 13



Plus de 4000 retweets en 5 minutes

Analyse du texte  
Analyse qualité expéditeur  
Systèmes automatisés ....

En fait des Hackers Syrien

136 Milliards \$ de perte en quelques secondes



# Une entreprise de télécommunications

---

- Les consommations des utilisateurs sur 3 ans

Number	Name	Phone	City	Plan	Avg. 3m Profit in \$
1	Nicholson Jack	647 224 8984	Paris	2y	12,00
3	Streep Meryl	647 231 3938	London	3y	189,45
4	De Niro Robert	633 345 8799	New York	3y	77,10
6	Pacino Al	654 478 7488	Singapore	3y	369,00
7	Day-Lewis Daniel	688 666 3431	Dellhi	3y	131,00
8	Hoffman Dustin	655 879 9963	Tokyo	2y	459,37
11	Monroe Marilyn	613 742 7361	Beijing	3y	830,00
12	Hopkins Anthony	638 378 6380	Cairo	3y	38,78
15	Newman Paul	633 789 7892	Jakarta	3y	299,29
17	Washington Denzel	624 798 2343	Bogota	1y	236,06
18	Winslet Kate	656 980 8793	Hanoi	3y	50,18
20	Penn Sean	645 892 8921	Santiago	3y	628,01
21	Blanchett Cate	635 891 1890	Berlin	3y	33,79
22	DiCaprio Leonardo	643 909 8918	Nairobi	3y	8,00
24	Brando Marlon	627 713 1053	Los Angeles	3y	26,23
26	Hanks Tom	667 017 6390	Montpellier	2y	89,11
28	Bridges Jeff	698 382 8614	Toronto	3y	92,75
31	Crowe Russel	689 139 4947	Munich	3y	1 044,48
33	Kidman Nicole	674 270 7824	Tokyo	3y	0,96



# Une entreprise de télécommunications

---

- Un problème de rentabilité
  - Il faut supprimer les utilisateurs non rentables
  - Lesquels faut il garder ?
  - Quel message donner aux autres pour les conserver ?
- Hypothèses :
  - les utilisateurs sont indépendants
  - Pas de particularité sur la distribution des valeurs de profit



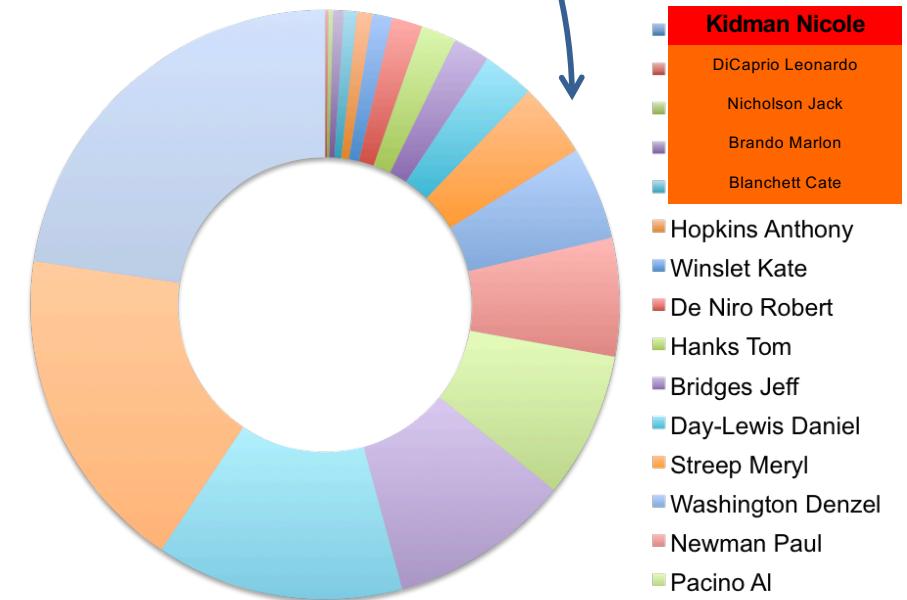
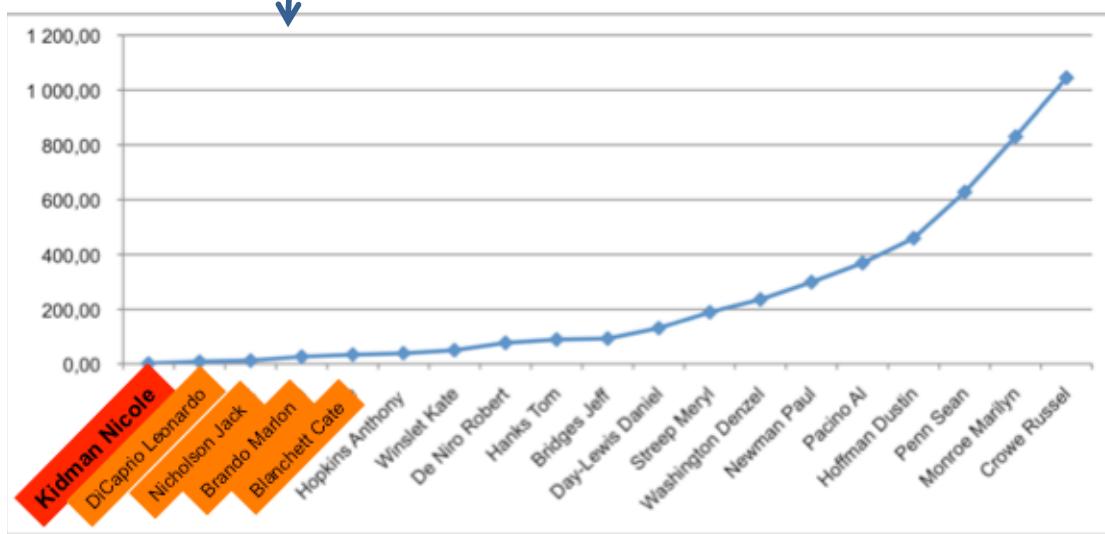
# Une approche classique - 1

- Un aperçu de la distribution



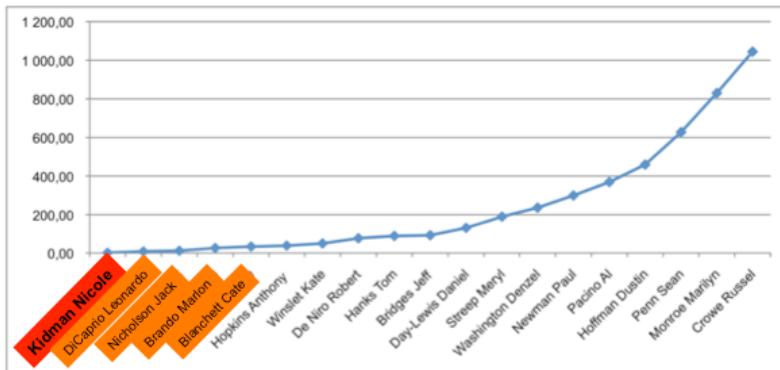
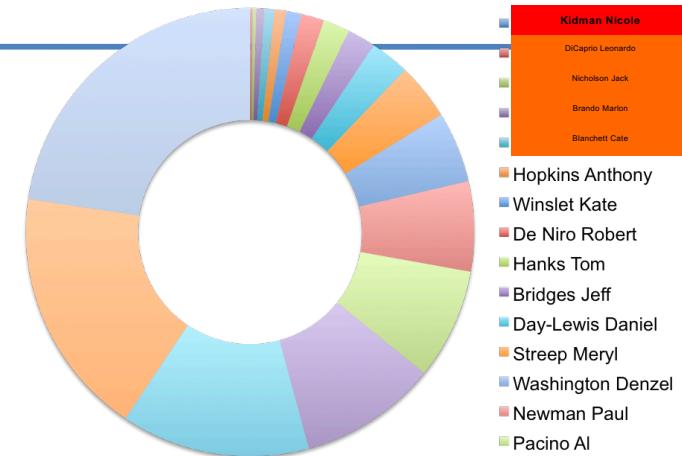
# Une approche classique - 2

Number	Name	Phone	City	Plan	Avg. 3m Profit in \$
33	Kidman Nicole	674 270 7824	Tokyo	3y	0,96
22	DiCaprio Leonardo	643 909 8918	Nairobi	3y	8,00
1	Nicholson Jack	647 224 8984	Paris	2y	12,00
24	Brando Marlon	627 713 1053	Los Angeles	3y	26,23
21	Blanchett Cate	635 891 1890	Berlin	3y	33,79
12	Hopkins Anthony	638 378 6380	Cairo	3y	38,78
18	Winslet Kate	656 980 8793	Hanoi	3y	50,18
4	De Niro Robert	633 345 8799	New York	3y	77,10
26	Hanks Tom	667 017 6390	Montpellier	2y	89,11
28	Bridges Jeff	698 382 8614	Toronto	3y	92,75
7	Day-Lewis Daniel	688 666 3431	Delhi	3y	131,00
3	Streep Meryl	647 231 3938	London	3y	189,45
17	Washington Denzel	624 798 2343	Bogota	1y	236,06
15	Newman Paul	633 789 7892	Jakarta	3y	299,29
6	Pacino Al	654 478 7488	Singapore	3y	369,00
8	Hoffman Dustin	655 879 9963	Tokyo	2y	459,37
20	Penn Sean	645 892 8921	Santiago	3y	628,01
11	Monroe Marilyn	613 742 7361	Beijing	3y	830,00
31	Crowe Russel	689 139 4947	Munich	3y	1 044,48



# Conclusions

33	Kidman Nicole	<b>0,96</b>
22	DiCaprio Leonardo	<b>8,00</b>
1	Nicholson Jack	<b>12,00</b>
24	Brando Marlon	<b>26,23</b>
21	Blanchett Cate	<b>33,79</b>
12	Hopkins Anthony	<b>38,78</b>



Clients à ne pas retenir  
6 sur 19  
Gain : **119,76 \$**

# Une intuition...

---

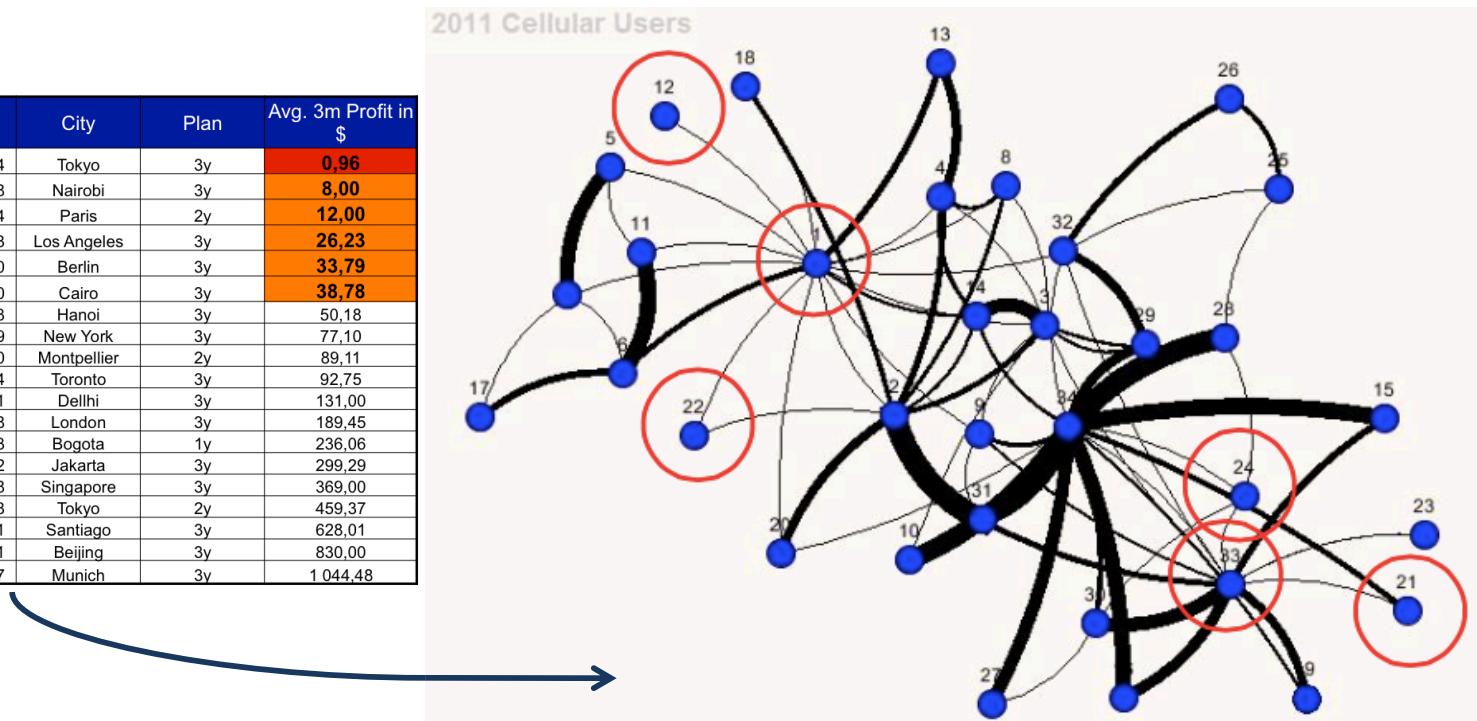
- Big data ... beaucoup de données ?
- et si il y avait d'autres données ?
  - *Data Linking* et intégration



# Données additionnelles

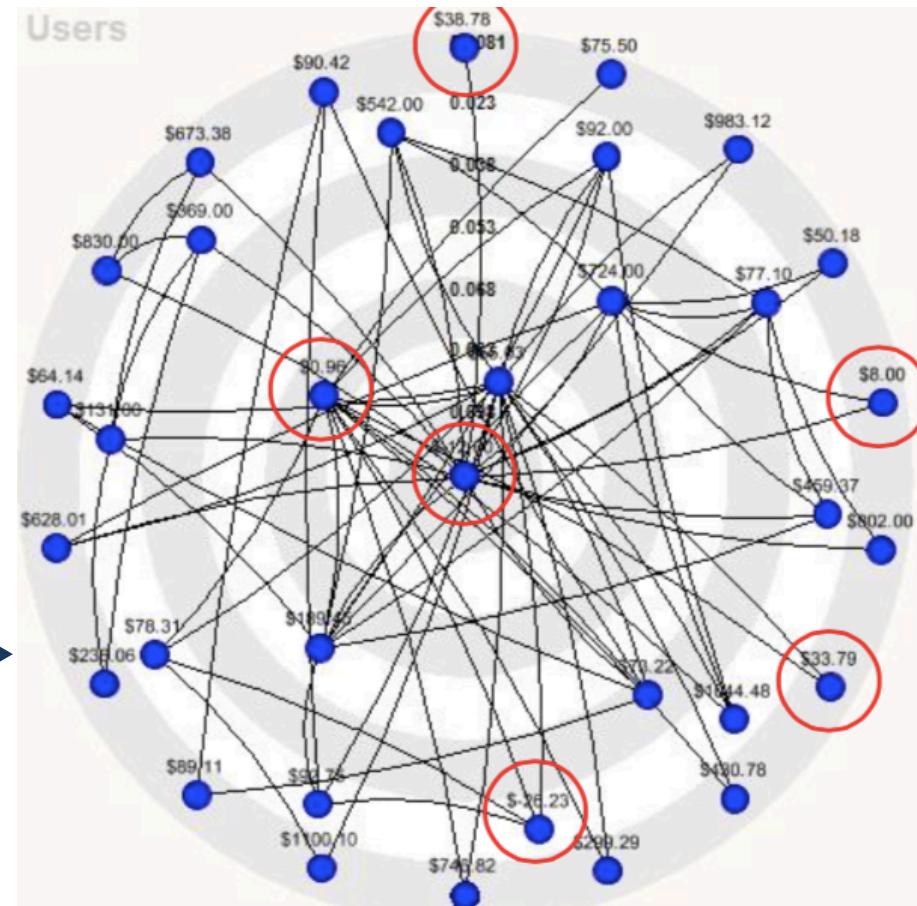
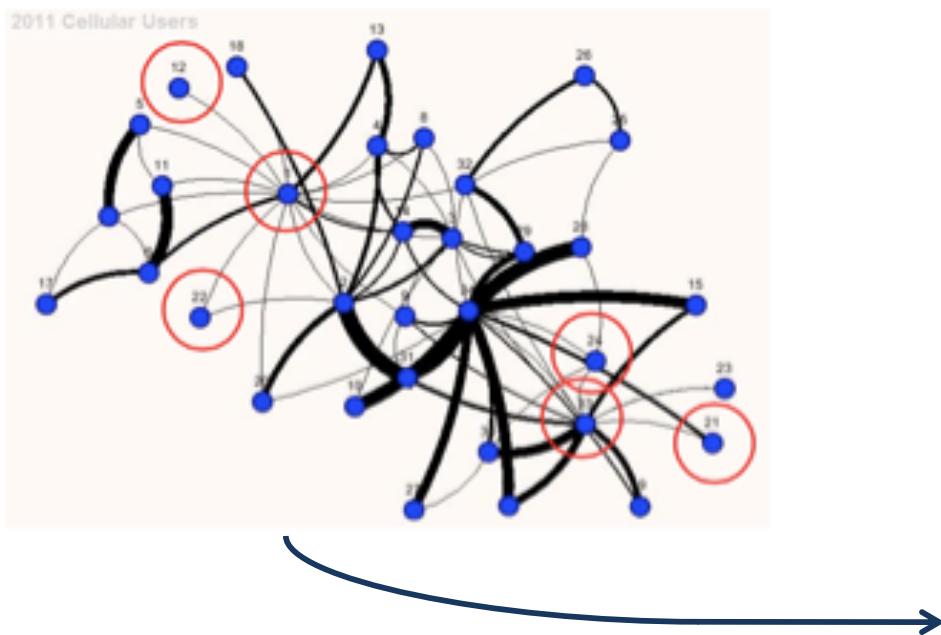
- « Inter-call network » avec les fréquences
- Ceux qui sont connectées avec les 19 personnes

Number	Name	Phone	City	Plan	Avg. 3m Profit in \$
33	Kidman Nicole	674 270 7824	Tokyo	3y	0,96
22	DiCaprio Leonardo	643 909 8918	Nairobi	3y	8,00
1	Nicholson Jack	647 224 8984	Paris	2y	12,00
24	Brando Marlon	627 713 1053	Los Angeles	3y	26,23
21	Blanchett Cate	635 891 1890	Berlin	3y	33,79
12	Hopkins Anthony	638 378 6380	Cairo	3y	38,78
18	Winslet Kate	656 980 8793	Hanoi	3y	50,18
4	De Niro Robert	633 345 8799	New York	3y	77,10
26	Hanks Tom	667 017 6390	Montpellier	2y	89,11
28	Bridges Jeff	698 382 8614	Toronto	3y	92,75
7	Day-Lewis Daniel	688 666 3431	Delhi	3y	131,00
3	Streep Meryl	647 231 3938	London	3y	189,45
17	Washington Denzel	624 798 2343	Bogota	1y	236,06
15	Newman Paul	633 789 7892	Jakarta	3y	299,29
6	Pacino Al	654 478 7488	Singapore	3y	369,00
8	Hoffman Dustin	655 879 9963	Tokyo	2y	459,37
20	Penn Sean	645 892 8921	Santiago	3y	628,01
11	Monroe Marilyn	613 742 7361	Beijing	3y	830,00
31	Crowe Russel	689 139 4947	Munich	3y	1 044,48



# Données additionnelles

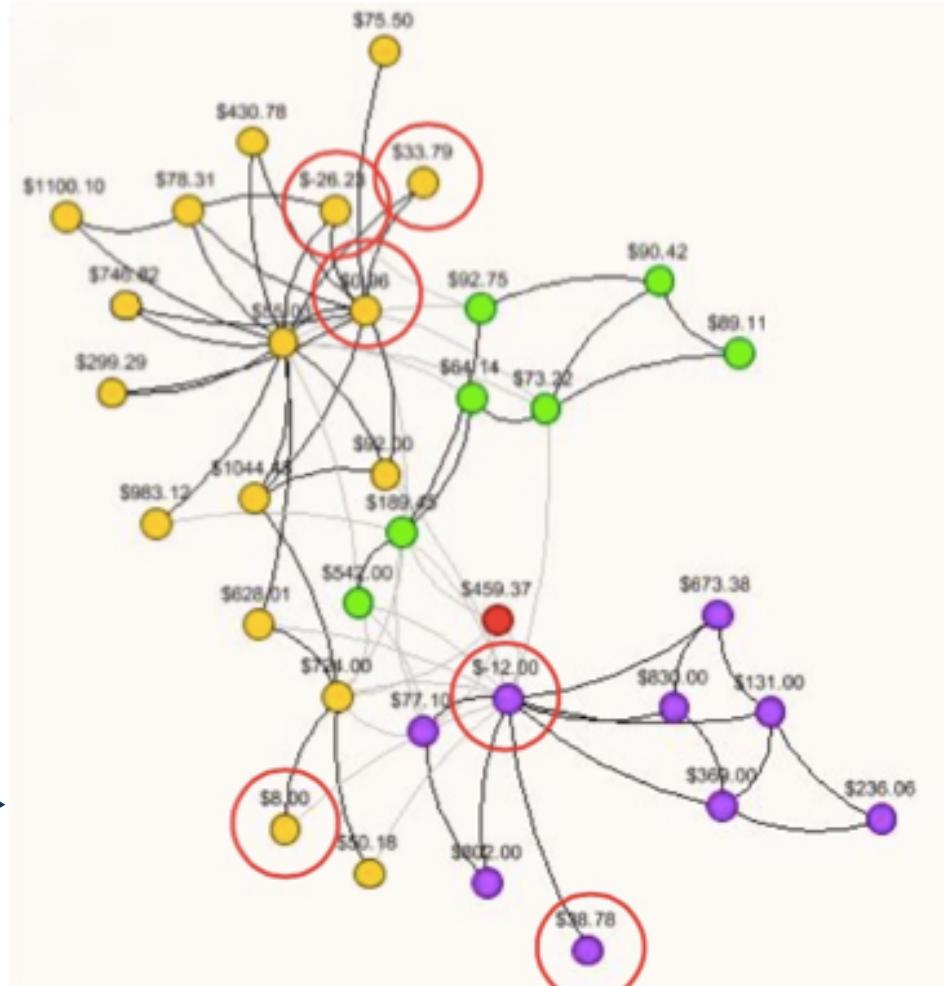
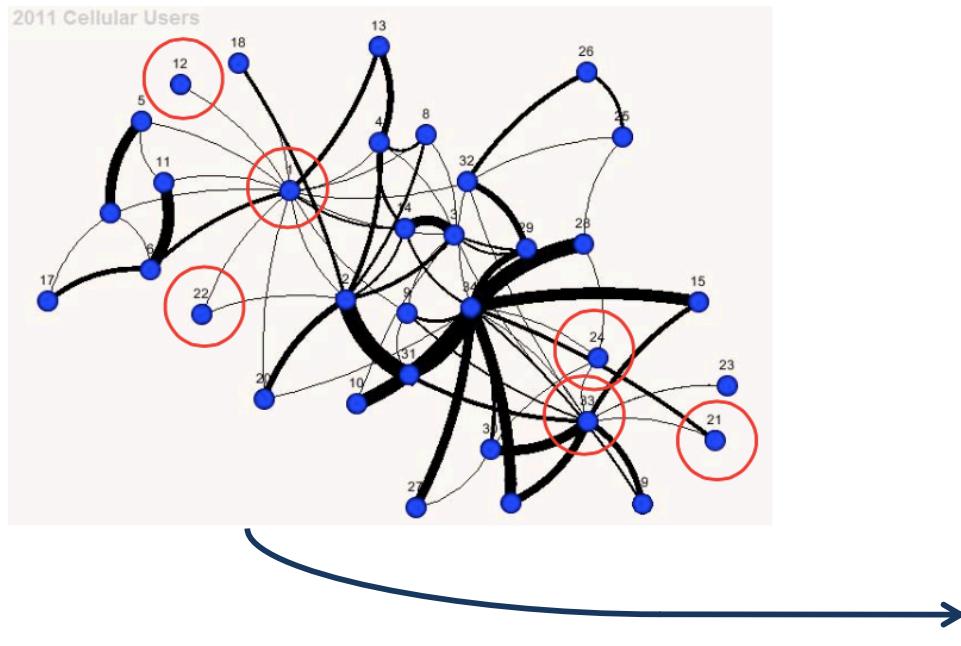
- Algorithmes de détection de communautés (*global community detection*)



Application du PageRank de Google <sup>26</sup>

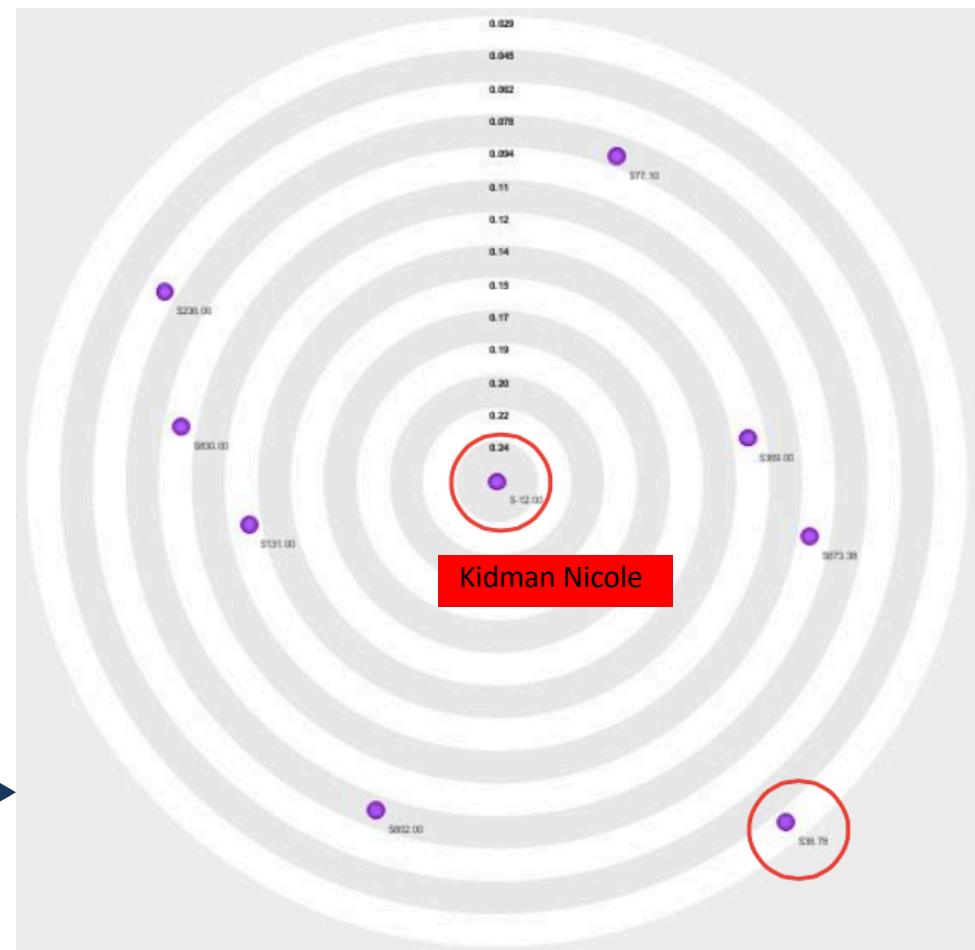
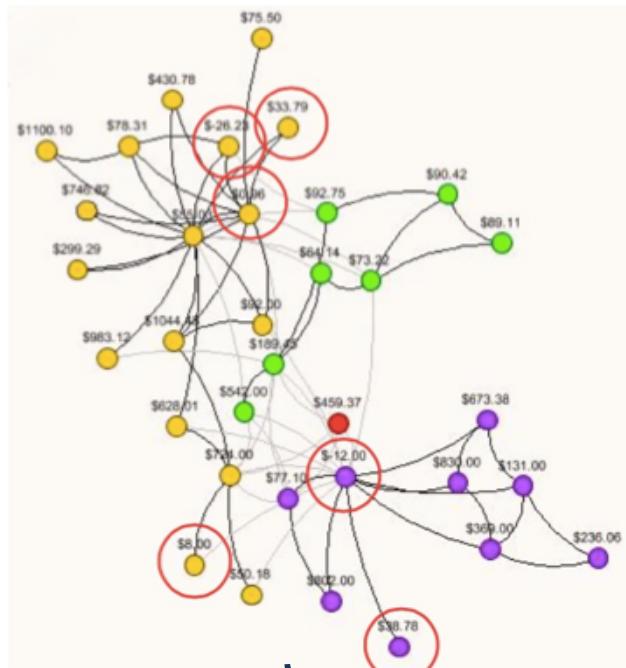
# Données additionnelles

- Algorithme de détection de communautés (*local community mining*)



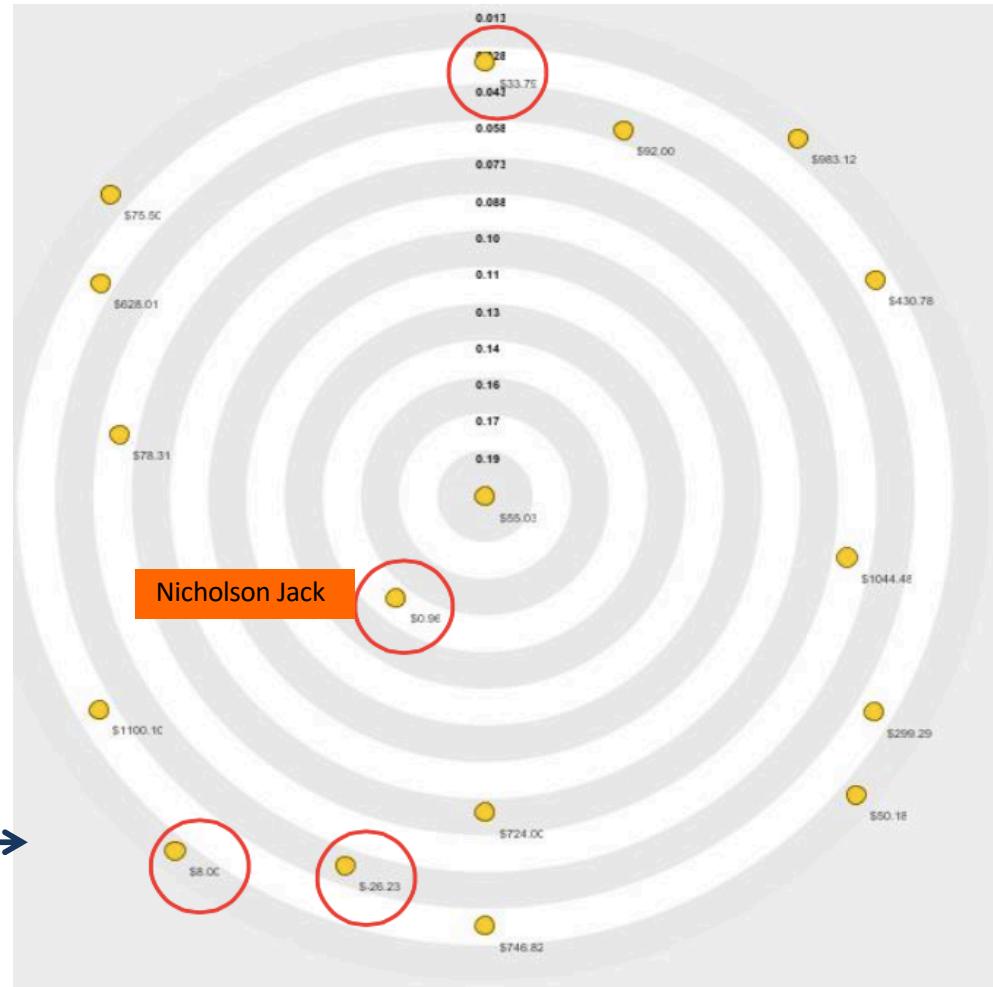
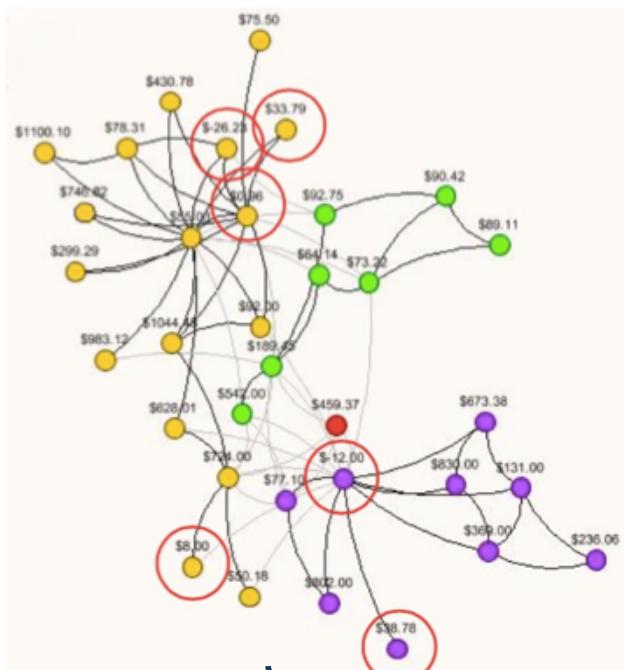
# Données additionnelles

- Centralité par Communauté (Nicole Kidman)

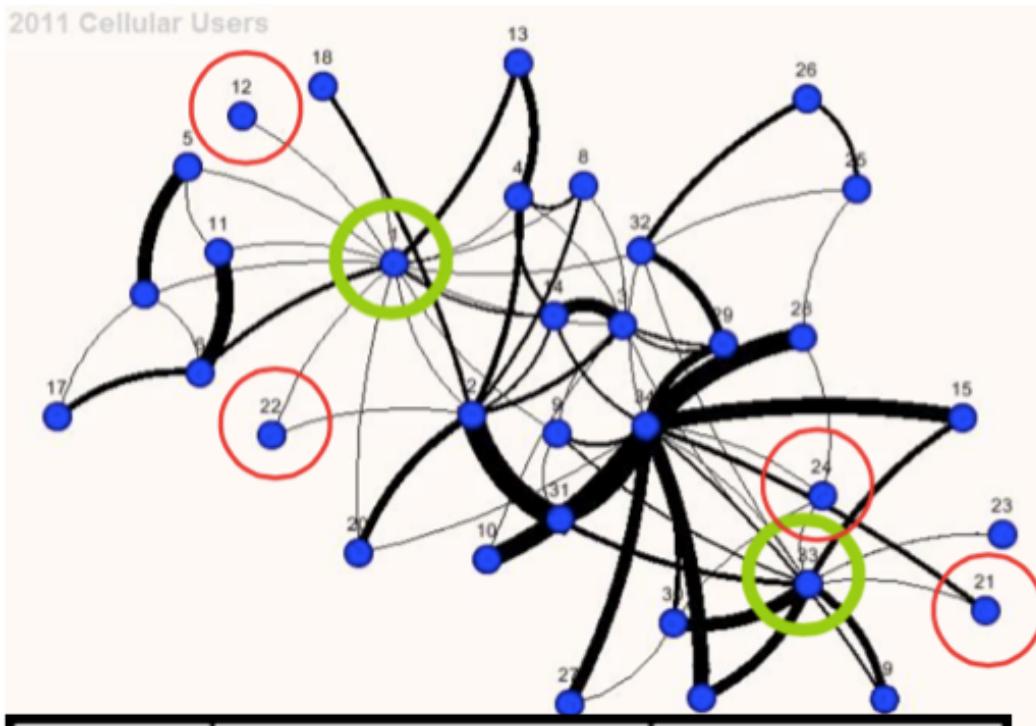


# Données additionnelles

- Centralité par Communauté (Jack Nicholson)



# Autres conclusions



33	Kidman Nicole	0,96
22	DiCaprio Leonardo	8,00
1	Nicholson Jack	12,00
24	Brando Marlon	26,23
21	Blanchett Cate	33,79
12	Hopkins Anthony	38,78

Risque de perte :

Nicole Kidman : 3145,32 \$ (0,96 \$)

Jack Nicholson : 6324,14 \$ (8 \$)

Exploiter des données additionnelles et des techniques d'analyses sophistiquées peuvent offrir de nouvelles perspectives

# A l'origine ...

---

## Application-Controlled Demand Paging for Out-of-Core Visualization

Michael Cox

MRJ/NASA Ames Research Center

Microcomputer Research Labs, Intel Corporation

[mbc@nas.nasa.gov](mailto:mbc@nas.nasa.gov)

David Ellsworth

MRJ/NASA Ames Research Center

[ellsworth@nas.nasa.gov](mailto:ellsworth@nas.nasa.gov)

### Abstract

In the area of scientific visualization, input data sets are often very large. In visualization of Computational Fluid Dynamics (CFD) in particular, input data sets today can surpass 100 Gbytes, and are expected to scale with the ability of supercomputers to generate them. Some visualization tools already partition large data sets into segments, and load appropriate segments as they are needed. However, this does not remove the problem for two reasons: 1) there are data sets

### 1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources. This *write-a-check* algorithm has two drawbacks. First, if visualization algorithms and tools are worth developing, then they are worth deploying to more

### Publication of:

- Conference

VIS97 IEEE Visualization '97 Conference

Phoenix, AZ, USA — **October 18 - 24, 1997**

IEEE Computer Society Press Los Alamitos, CA, USA ©1997



# Le Big Data s'affiche...

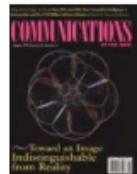


## Visually exploring gigabyte data sets in real time

Full Text: [Get this Article](#)

Authors: [Steve Bryson](#) NASA Ames Research Center, Moffett Field, CA  
[David Kenwright](#) NASA Ames Research Center, Moffett Field, CA  
[Michael Cox](#) NASA Ames Research Center, Moffett Field, CA  
[David Ellsworth](#) NASA Ames Research Center, Moffett Field, CA  
[Robert Haimes](#) Massachusetts Institute of Technology, Cambridge

### Published in:



- Magazine  
Communications of the ACM [Homepage archive](#)
- Volume 42 Issue 8, Aug. 1999  
Pages 82-90  
ACM New York, NY, USA  
[table of contents](#) doi:>[10.1145/310930.310977](https://doi.org/10.1145/310930.310977)

## Introduction

David N. Kenwright

MRJ Technology Solutions  
NASA Ames Research Center

davidk@nas.nasa.gov

Exploring Gigabyte Datasets in Real-Time - Introduction

1 - 1 - 1

## Big Data

There are two distinct types of "big data"

- big data collections
- big data objects

Big data collections are aggregates of many datasets  
(e.g., data warehouses, airline reservation systems)

Big data objects are single datasets from computer simulations  
(e.g., computational fluid dynamics, structural analysis)

Exploring Gigabyte Datasets in Real-Time - Introduction

1 - 1 - 2



# Numéro spécial dans Nature

[nature](#) International weekly journal of science

Journal home > Archive > Editorial > Full Text

**Journal content**

- [+ Journal home](#)
- [+ Advance online publication](#)
- [+ Current issue](#)
- [+ Nature News](#)
- [+ Archive](#)
- [+ Supplements](#)
- [+ Web focuses](#)
- [+ Podcasts](#)
- [+ Videos](#)
- [+ News Specials](#)

**Editorial**

*Nature* 455, 1 (4 September 2008) | doi:10.1038/455001a; Published online 3 September 2008

**Community cleverness required**

**Researchers need to adapt their institutions and practices in response to torrents of new data — and need to complement smart science with smart searching.**

The Internet search firm Google was incorporated just 10 years ago this week. Going from a collection of donated servers housed under a desk to a global network of dedicated data centres processing information by the petabyte, Google's growth mirrors that of the production and exploration of data in research. All of which makes this an apt moment for this special issue of *Nature*, which examines what big data sets mean for contemporary science.

'Big', of course, is a moving target. The portability of the tens of gigabytes we carry around on USB sticks would have seemed like

**subscribe to nature**

**FULL TEXT**

- [+ Previous | Next +](#)
- [+ Table of contents](#)
- [Download PDF](#)
- [View interactive PDF in ReadCube](#)
- [Share this article](#)
- [CrossRef lists 33 articles citing this article](#)
- [Scopus lists 34 articles citing this article](#)



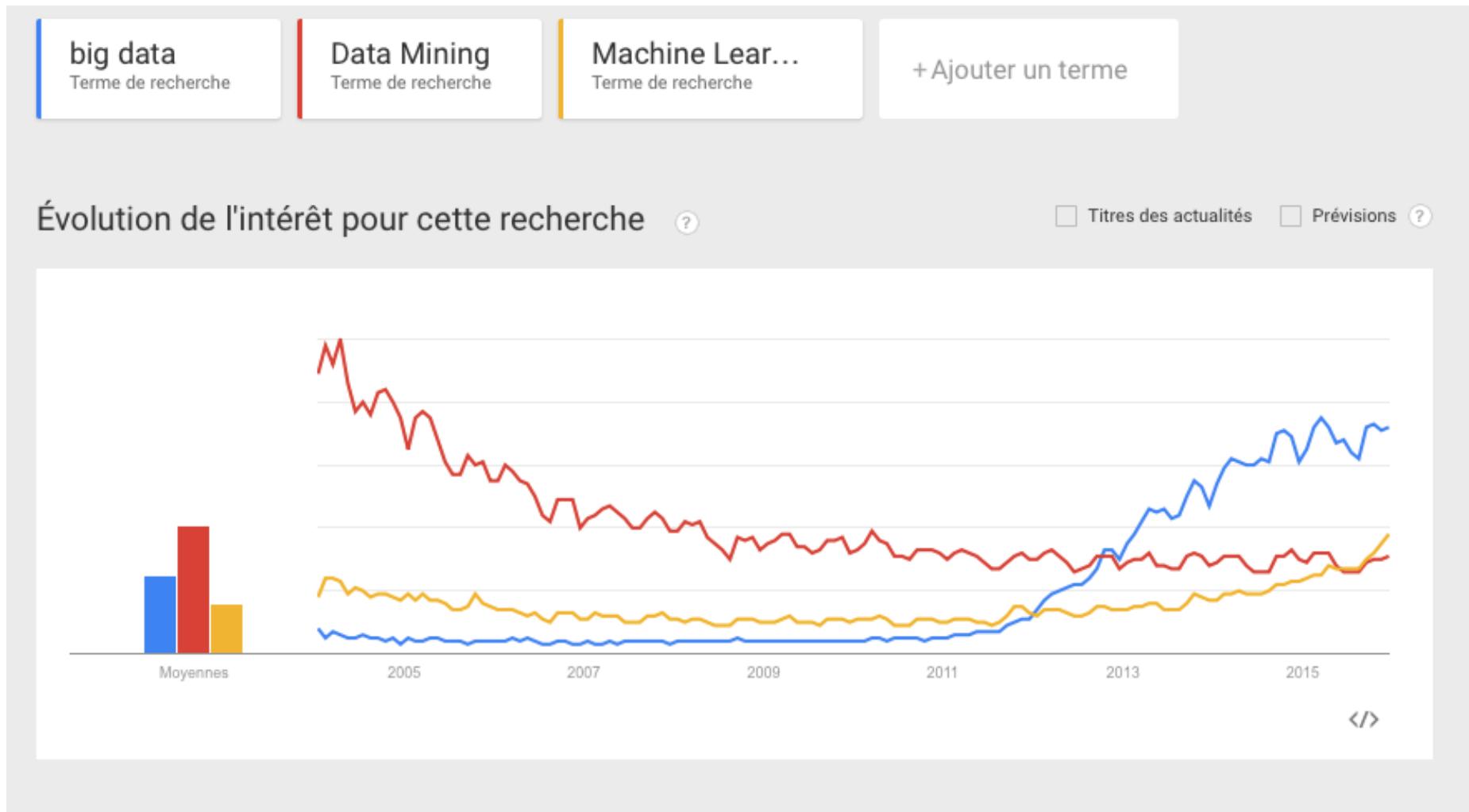
# Google Trends



# Big Data vs Data Mining



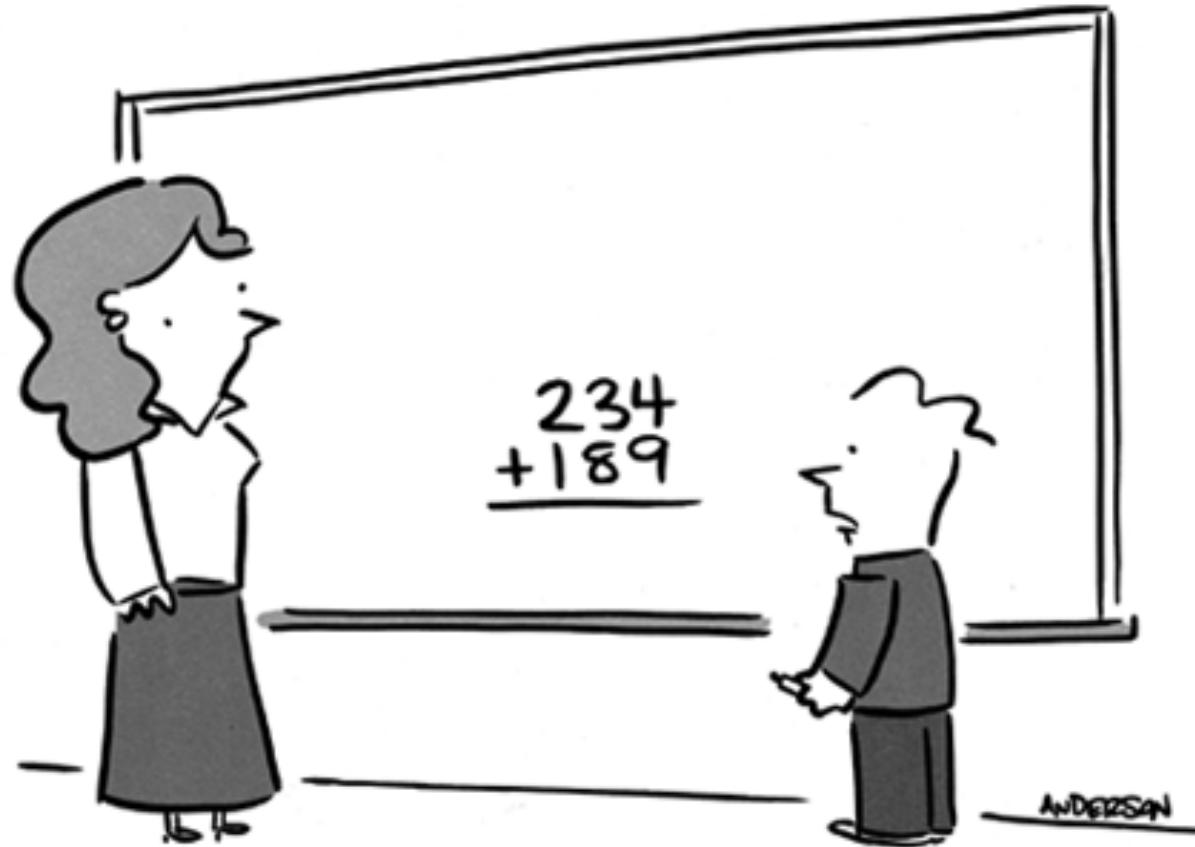
# Big Data vs Data Mining vs Machine Learning



# Aujourd’hui le Big Data est partout

© MARK ANDERSON

WWW.ANDERTOONS.COM



"Does this count as big data?"

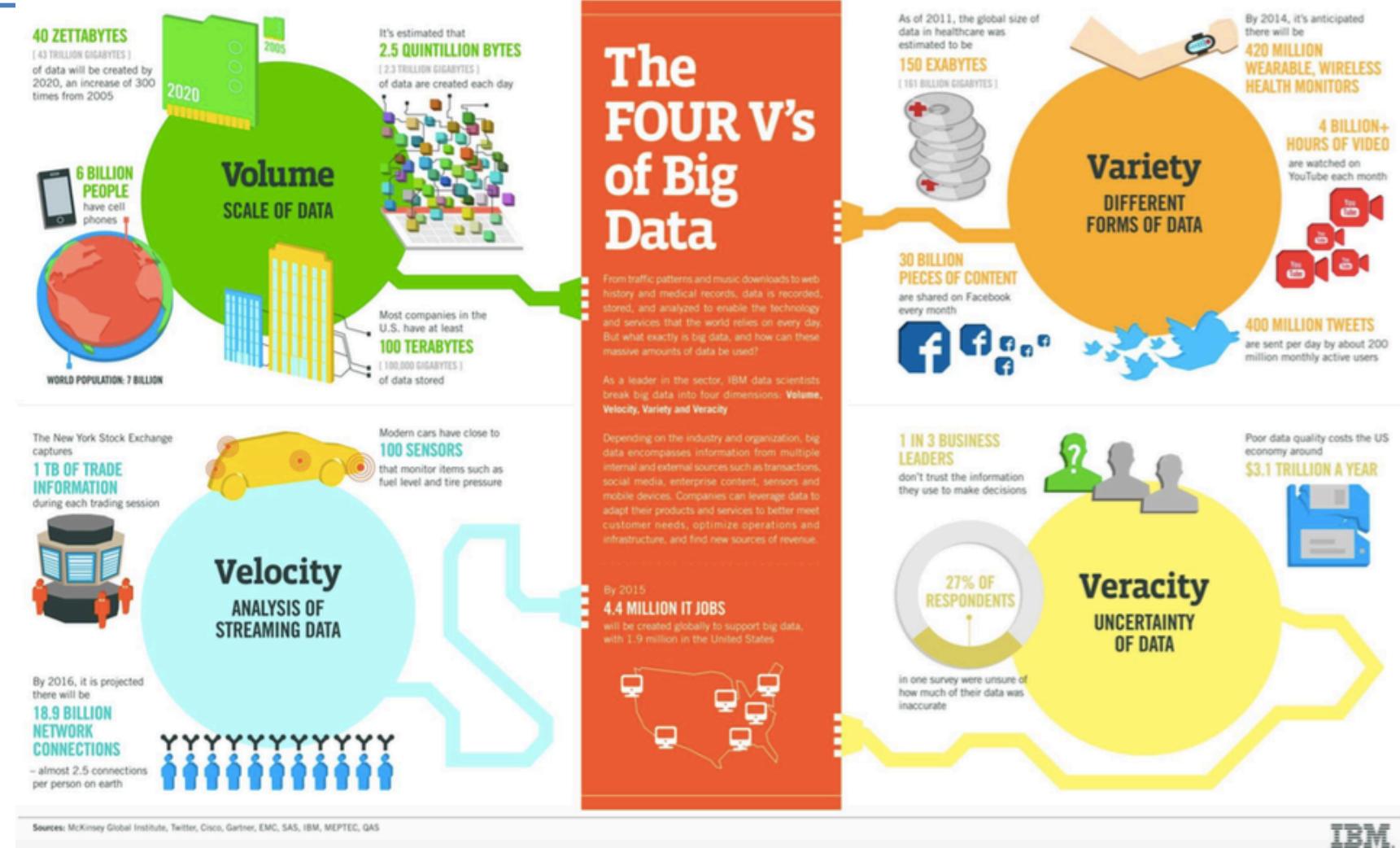
LET'S SOLVE THIS PROBLEM BY  
USING THE BIG DATA NONE  
OF US HAVE THE SLIGHTEST  
IDEA WHAT TO DO WITH



© marketoonist.com

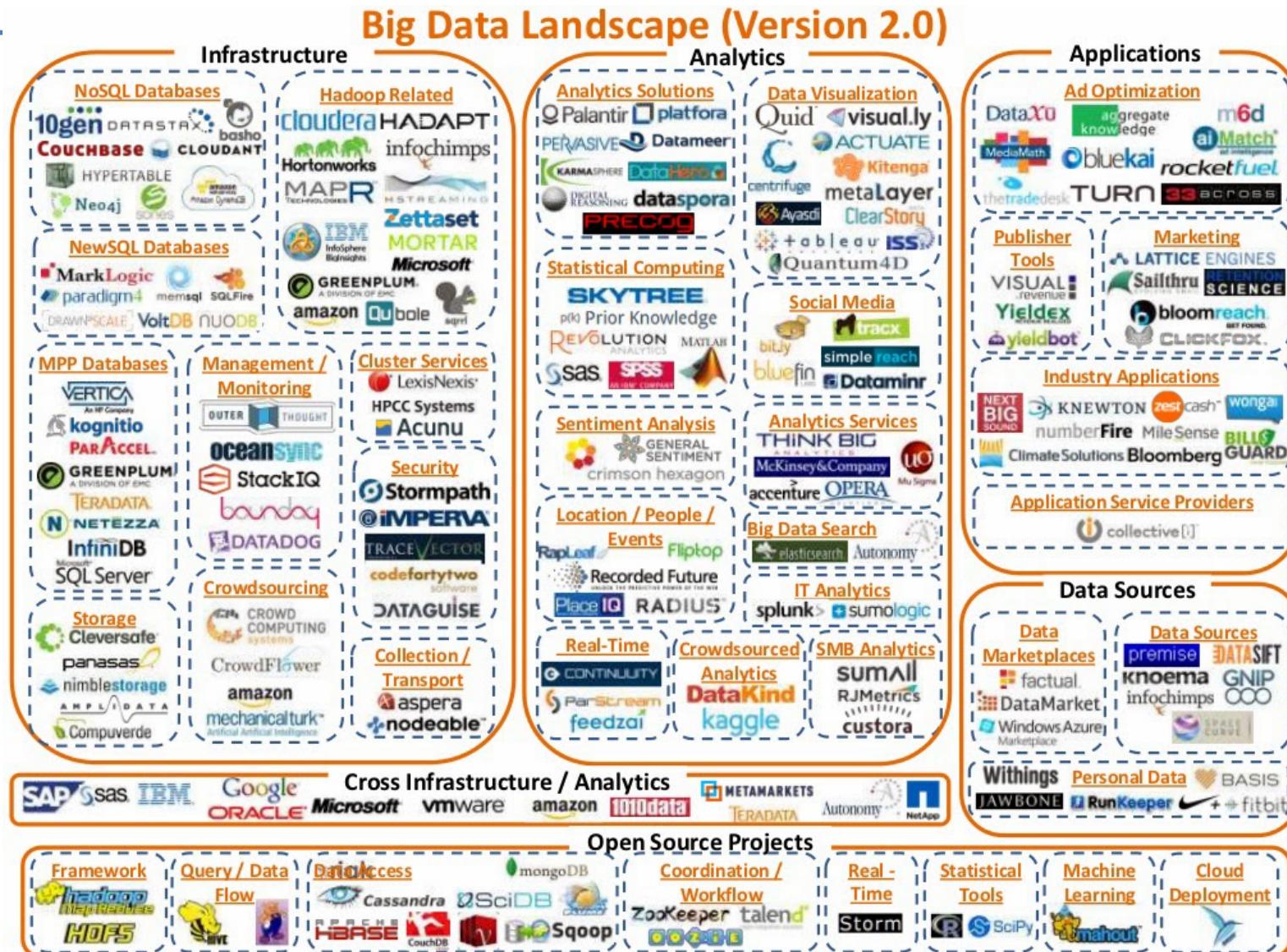


# The Vs or the dimensions of Big Data



IBM

# Des outils existent ...

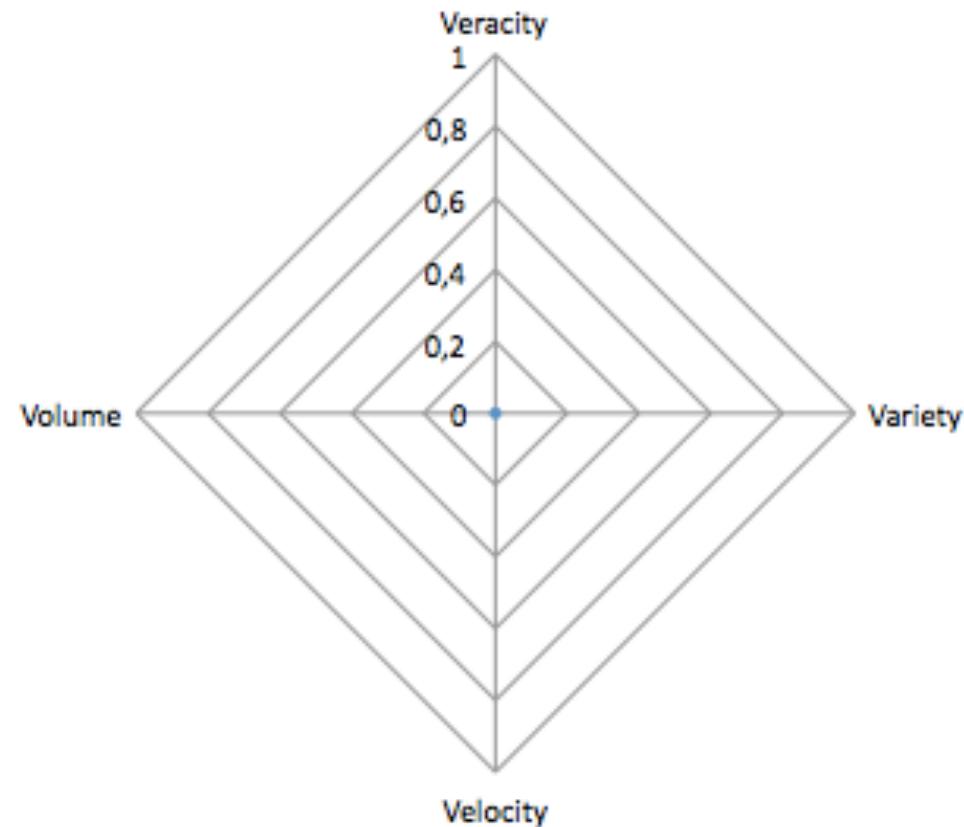


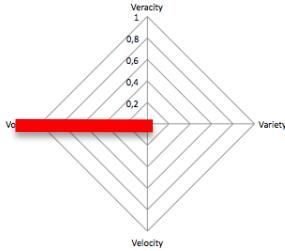
© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures



# Les 4 dimensions

---





# Dimension Volume

## Application-Controlled Demand Paging for Out-of-Core Visualization

Michael Cox  
MRJ/NASA Ames Research Center  
Microcomputer Research Labs, Intel Corporation  
[cmcb@nas.nasa.gov](mailto:cmcb@nas.nasa.gov)

David Ellsworth  
MRJ/NASA Ames Research Center  
[ellsworth@nas.nasa.gov](mailto:ellsworth@nas.nasa.gov)

### Abstract

In the area of scientific visualization, input data sets are often very large. In visualization of Computational Fluid Dynamics (CFD) in particular, input data sets today can surpass scale with the ability of Some visualization tools into segments, and load appropriate segments as they are needed. However, this does not remove the problem for two reasons: 1) there are data sets

100Gbytes

### 1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources. This *write-a-check* algorithm has two drawbacks. First, if visualization algorithms and tools are worth developing, then they are worth deploying to more

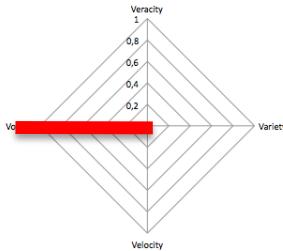
12 To  
625,55



Aujourd’hui

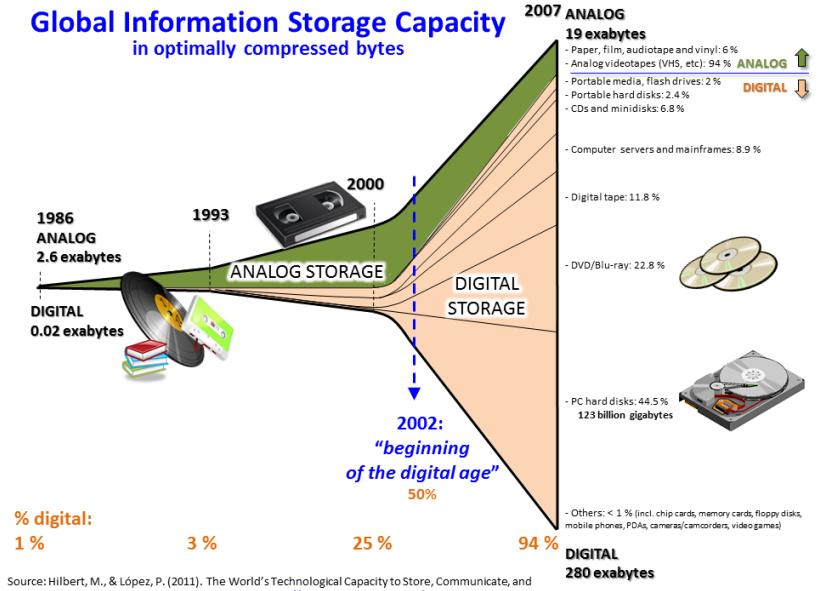


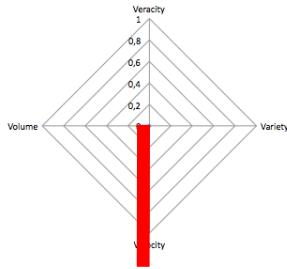
1997



# Dimension Volume

- De plus en plus de données
- Quelques challenges ?
  - Comment traiter des données de plus en plus volumineuses ?
  - Comment prendre en compte le fait qu'elles soient réparties (crowdsourcing, autre) ?
  - Comment accéder rapidement aux données ?



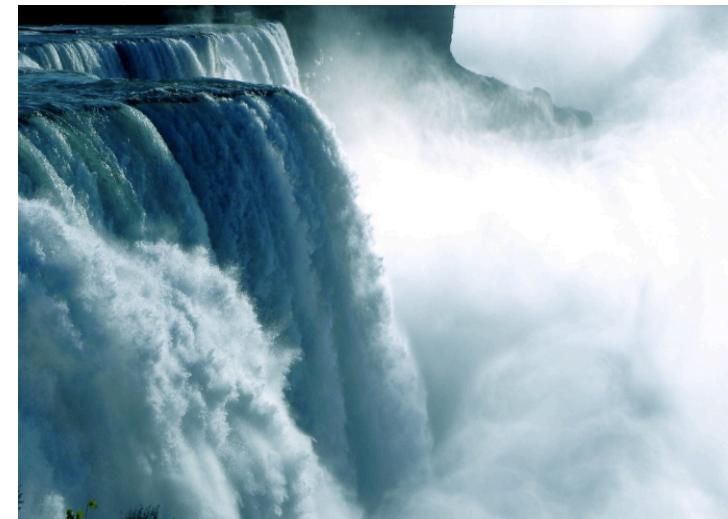


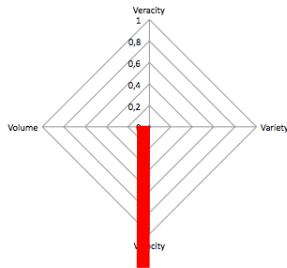
# Dimension *Velocity*

- Des données arrivant de plus en plus vite
- Quelques challenges ?
  - Comment supporter ces données (infrastructures) ?
  - Comment prendre en compte leur répartition ?
  - Comment gérer l'énergie des capteurs ?
  - Comment interroger/analyser ces données ?

## UNE VITESSE DE TRANSFERT DE DONNÉES DE 43 TBPS : UN RECORD DU MONDE ÉTABLI PAR DES CHERCHEURS DANOIS SUR FIBRE OPTIQUE

En utilisant un nouveau type de fibre optique, les chercheurs de l'Université technique du Danemark (DTU) ont transmis des données sur une seule fibre optique à une vitesse de 43 térabits par seconde (43 Tbps) et établi ainsi, un nouveau record du monde pour la transmission de données. Il bat le précédent record de 32 Tbps établi par les chercheurs du Karlsruhe Institute of Technology en Allemagne.





# Dimension Velocity

## Explosion of Data in Recent Years

- 3 Billion Telephone Calls in US each day
- 30 Billion email daily - 1 Billion SMS, IMs
- IP Network Traffic: up to 1 Billion packets per hour per router. Each Internet Service Provider has many (hundreds) of routers! 75000 tuples per second!
- AT&T collects 100 GBs of NetFlow data each day
  
- Scientific data: NASA EOS (Earth Observation System) observation satellites generate 350 GBs per day

3

## Explosion of Data in Recent Years

- eBay users worldwide trade more than \$1590 worth of goods every second
- eBay averages 1 Billion page views per day
- Yahoo (2002): 166 millions of visitors per day; 48 Gbs per hour of clickstream
  
- Compare to human scale data: « Only » 1 billion worldwide credit card transactions per month

4

on parle d'un parc de plus d'un million de serveurs. On estime à **plus de 200 000** le nombre de serveurs déployés chaque année ; à titre de comparaison, la taille du parc installé de toutes les entreprises du CAC40 est du même ordre de grandeur. Le nombre d'entreprises françaises possédant plus de 20000 serveurs se compte sur les doigts de la main... Cette puissance de feu permet à Amazon d'abriter le plus grand système de stockage au monde : S3, son service de stockage Objet contient ainsi **plus de 2 trillions d'objets**

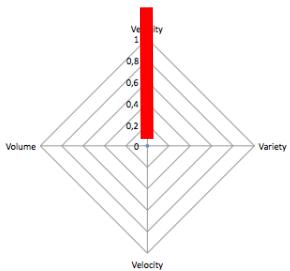
**Plus de 1 million de requêtes par seconde**

Mars 2014

Aujourd'hui ?

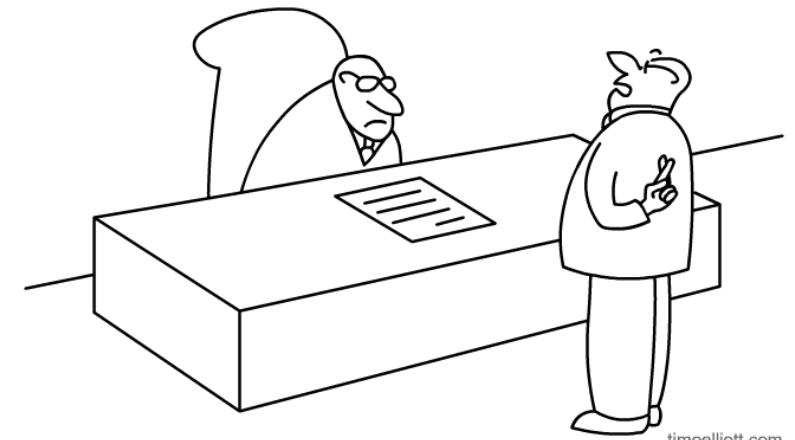


Cours 2011

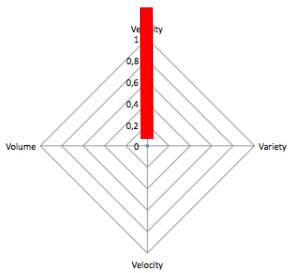


# Dimension *Veracity*

- Comment garantir qu'une donnée est valide ?
  - Une erreur dans un capteur ?
  - Un mauvais utilisateur ?
- Quelques challenges ?
  - Déetecter les erreurs ?
  - Traiter les erreurs ?
  - Cryptographie ?
  - Préservation de la vie privée ?



*"Yes sir, you can absolutely trust those numbers"*



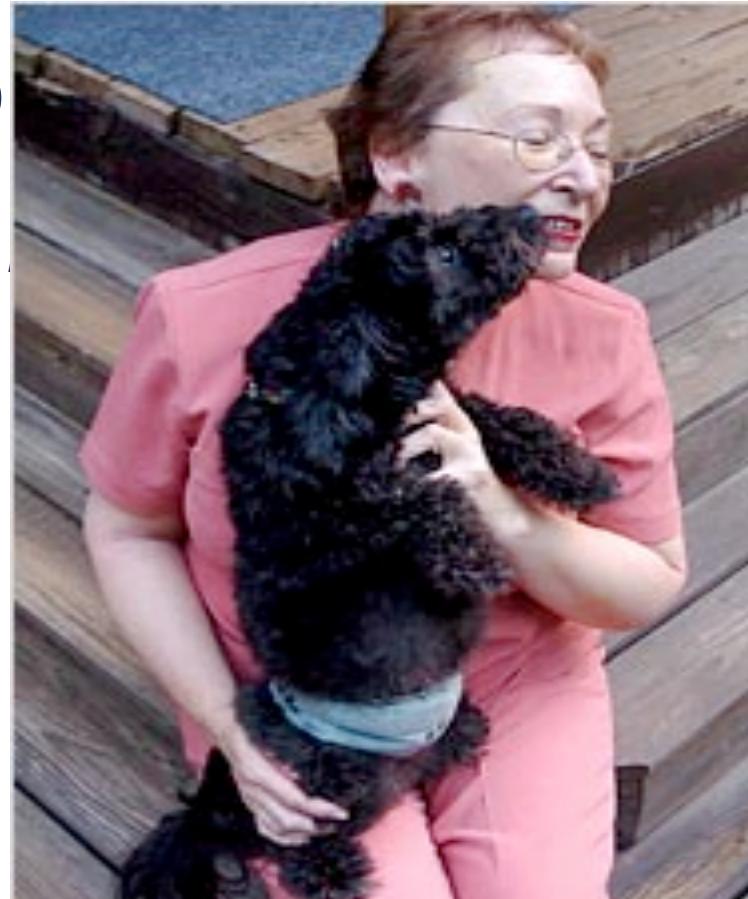
# Dimension Veracity

- L'Expérience d'AO

No. 4417749

[«homme célibataire de 60 ans», «

- Data linking



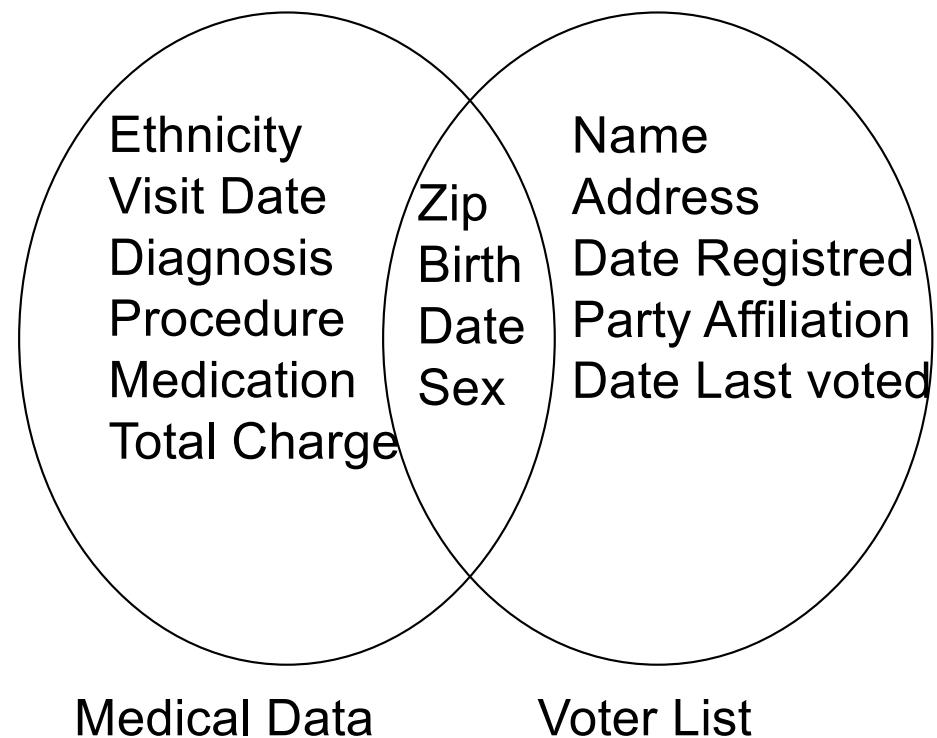
Thelma Arnold  
veuve de 62 ans  
Lilburn, Georgie

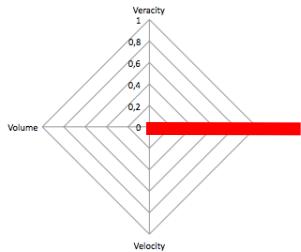
# Dimension Veracity

- Fichier anonymisé des soins de santé des fonctionnaires de l'état du Massachusetts mis en ligne (L. Sweeney, 1997)
- La liste électorale de Cambridge, MA (53 805 inscrits)

□ 69 % d'enregistrements uniques par rapport à code postal, date de naissance

## Dossier médical du gouverneur du Massachusetts





# Dimension Variety



**20**minutes.fr

leParisien.fr

**LE FIGARO**.fr

Le Monde.fr

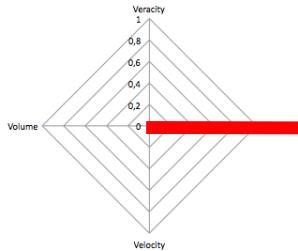
Liberation.fr

Slate.fr Rue89

Les Echos.fr

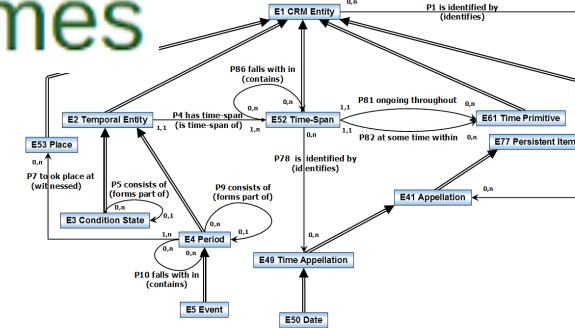
L'EXPRESS .fr  
MEDIAPART



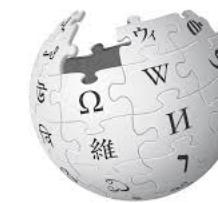


# Dimension Variety

- Quelques Challenges ?
  - Connaissances disponibles ?
  - Expertise ?
  - Linking Data
  - Open Data



BabelNet



WIKIPEDIA  
The Free Encyclopedia



# Une seule source de données

The collage includes:

- A large red banner at the top left with the text "Limitations of Big Data" and "Analytics".
- A photograph of a group of men holding a large ceremonial check from Netflix for \$1,000,000.
- A graph showing the RMS Error on Cinematch decreasing from approximately 0.95 to 0.85 over time, with a dashed line indicating further improvement.
- A screenshot of the Internet Movie Database (IMDb) website.
- A quote in the foreground: "A year into the competition, the Korbell team won the first Progress Prize with an 8.43% improvement. They reported more than 2000 hours of work in order to come up with the final combination of 107 algorithms that gave them this prize. And, they gave us the source code. We looked at the two underlying algorithms with the blend of these two reduced the error to 0.88. To put these algorithms to use, we had to work to overcome one limitation: the limitations that they were built to handle 100 million ratings, instead of the more than 5 billion that we have, and that they were not built to adapt as members added more ratings. But once we implemented it, it worked great."

# En réalité

---

- Plus d'une seule dimension !



# Linking Data ...

## Contrast to the Nate Silver Story

- In 2012, correctly predicted Obama winning re-election against Romney in all 50 states.
- Linked data instead of inventing a new algorithm
- Used Poll data and linked the data with more relevant information such as historical successes of those polls, Past 30 years of Election Databases for all states, ...
- Built a model using the linked data.



First Success  
Story of Big  
Data?



Michael Cosentino @cosentino Follow

For the Nate-haters, here's the 538 prediction and actual results side by side

Reply Retweet Favorite More

538 prediction      actual

RETWEETS 3,128 FAVORITES 723

8:58 PM - 6 Nov 2012 Flag media

Related headlines

- How did Nate Silver predict the US election? The Guardian @guardian
- Triumph of the Nerds: Nate Silver Wins in 50 States Mashable @mashable
- Obama's win a big vindication for Nate Silver, king of the quants CNET @CNET

# Quels sont les autres challenges ?

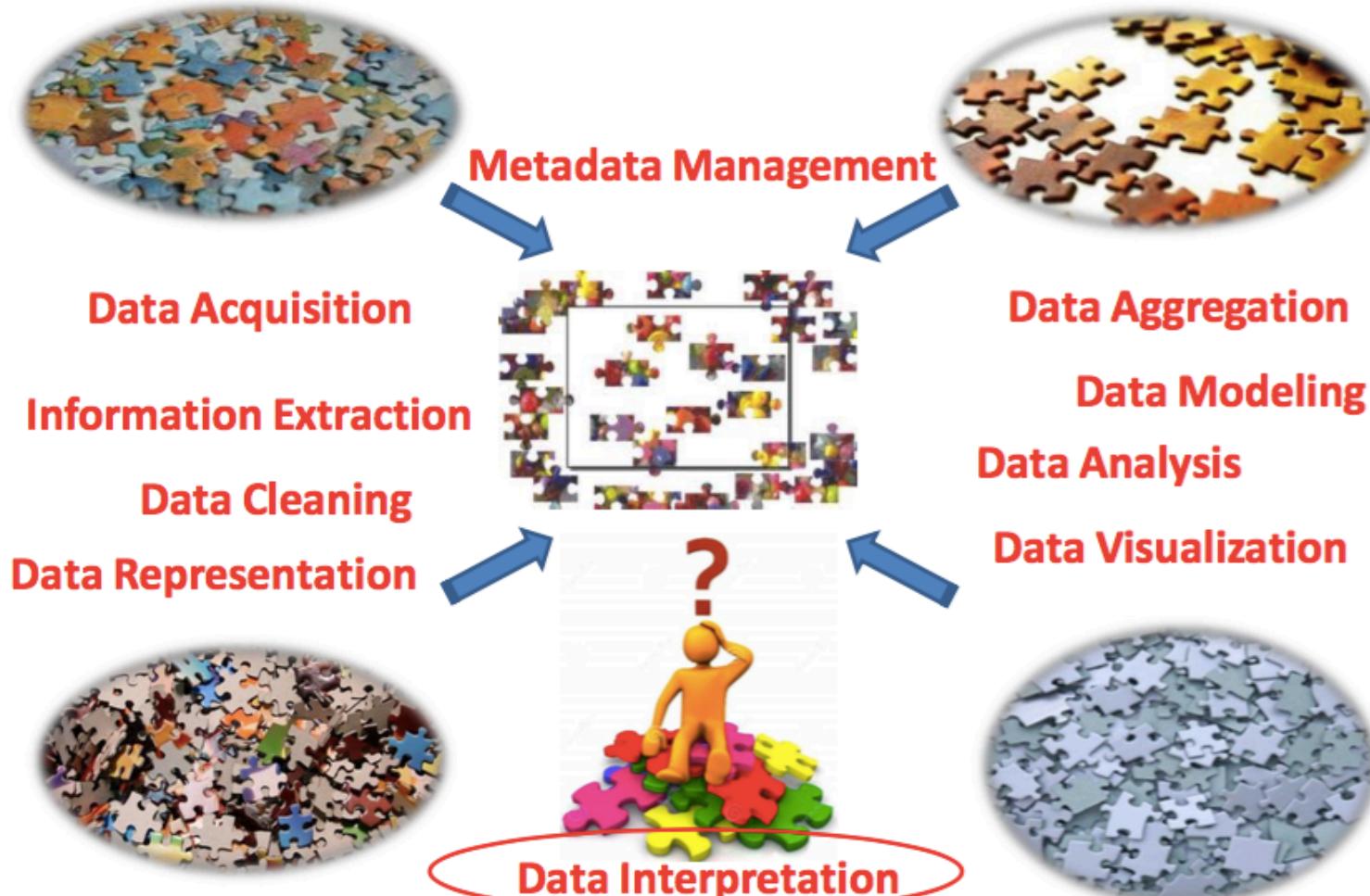
---

- Au moins deux grands challenges :
  - Intégration des données
  - Interprétation et analyse des données



# Intégration

Value is in Data Integration



# Comment analyser et interpréter ?

## VAlRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History

Isaac Cho, Wewnen Dou, Derek Xiaoyu Wang, Eric Sauda and William Ribarsky

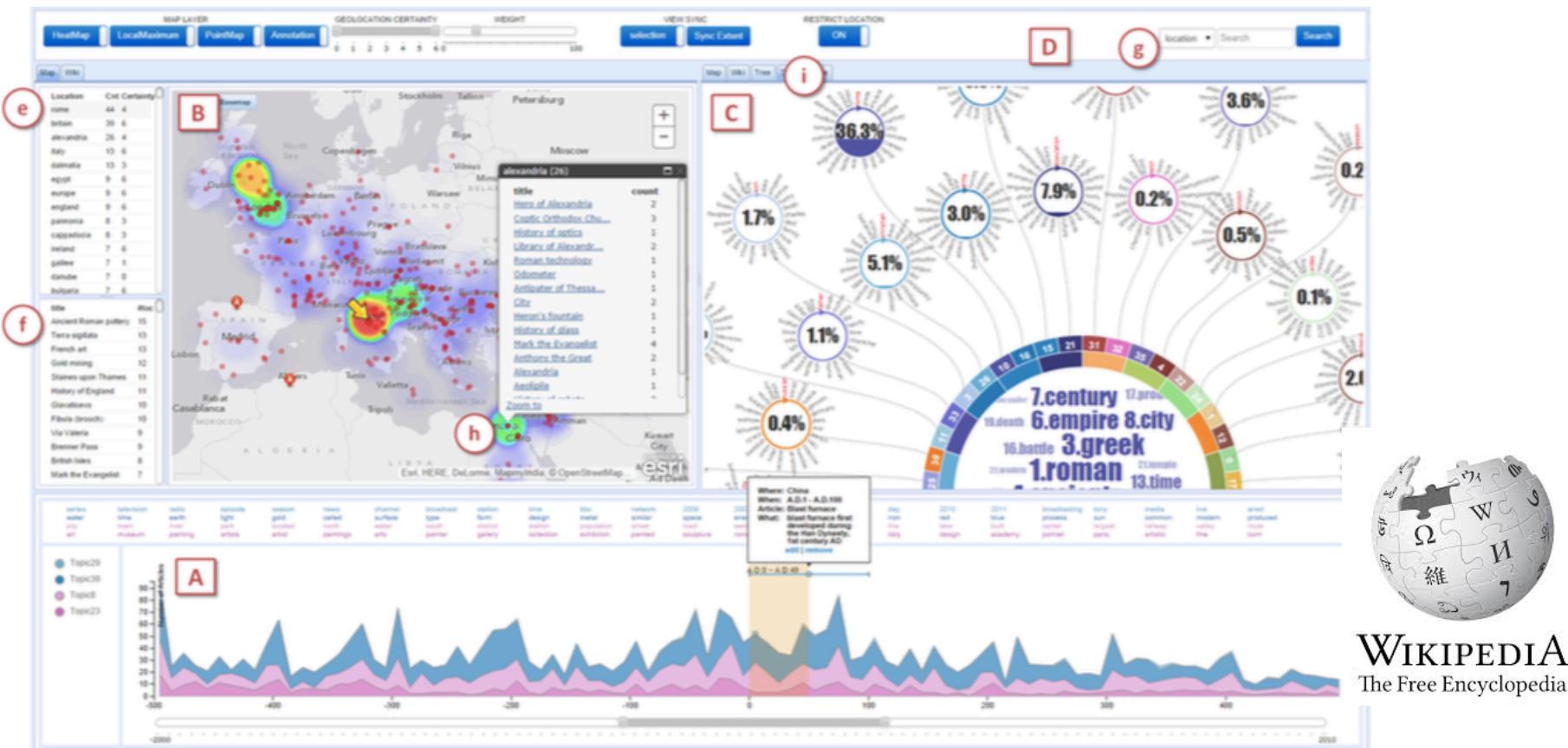


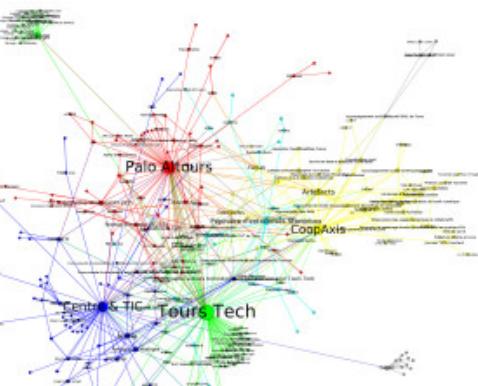
Fig. 1. Overview of VAlRoma Interface. The interface has three main views: Timeline view (A), Geographic view (B) and Topic view (C). A user-generated annotation is shown in the Timeline view.



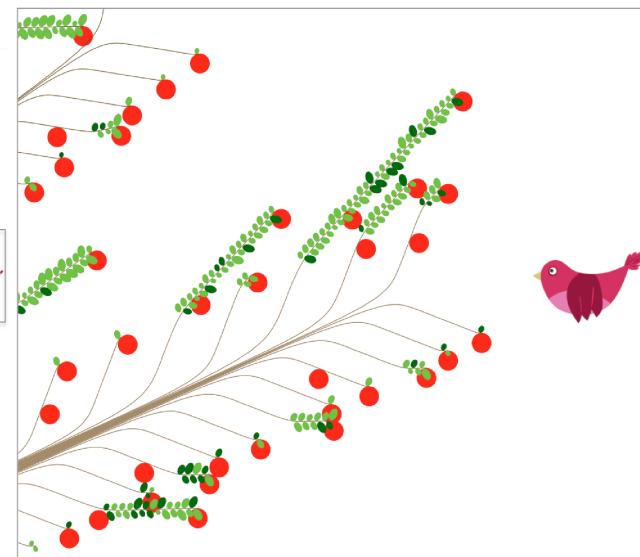
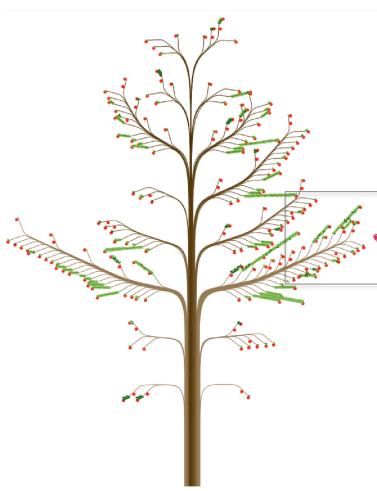
# Comment analyser et interpréter ?



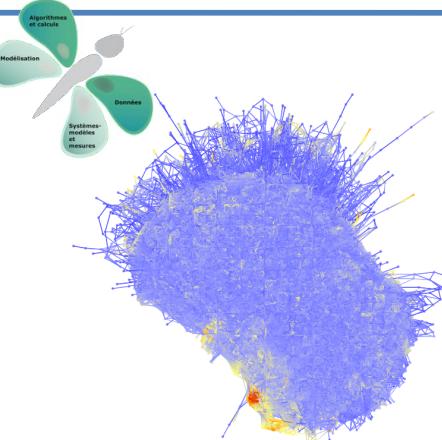
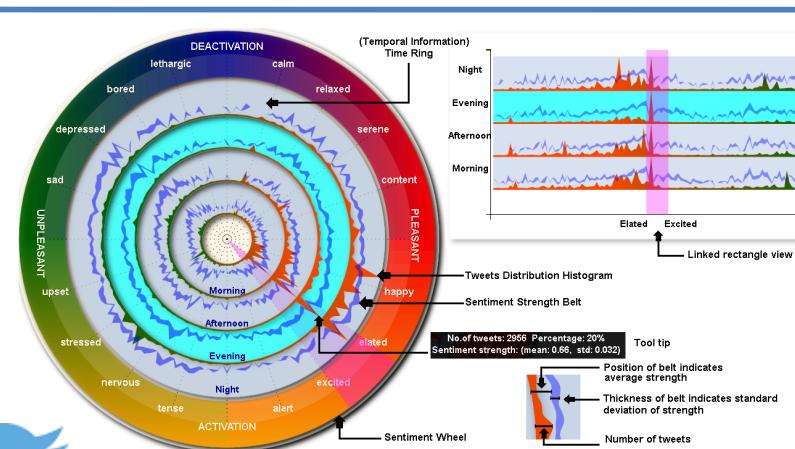
Analyse de sentiments dans les tweets



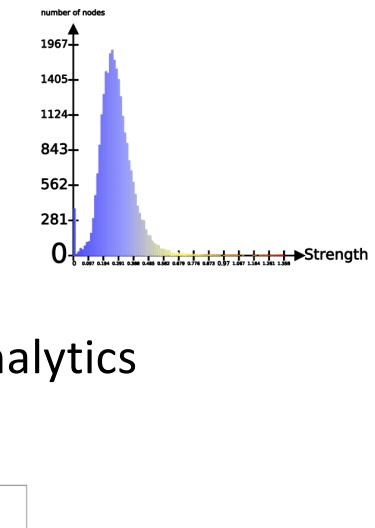
Graphes de communautés



De nouvelles abstractions

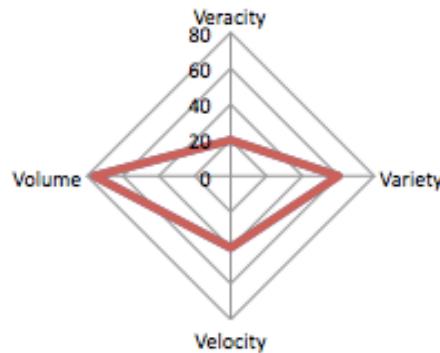


Visual Analytics

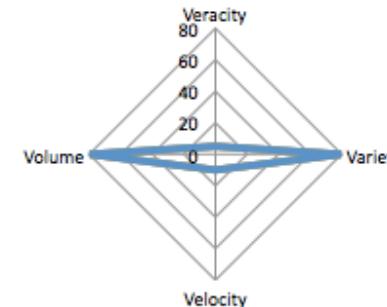


# La bonne question ?

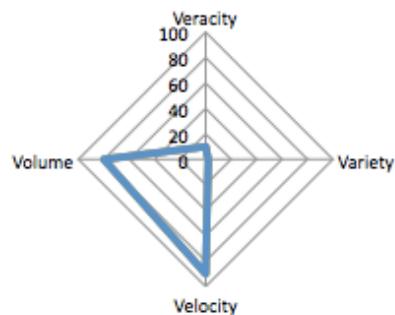
- Dans quelles dimensions se situe mon projet ?



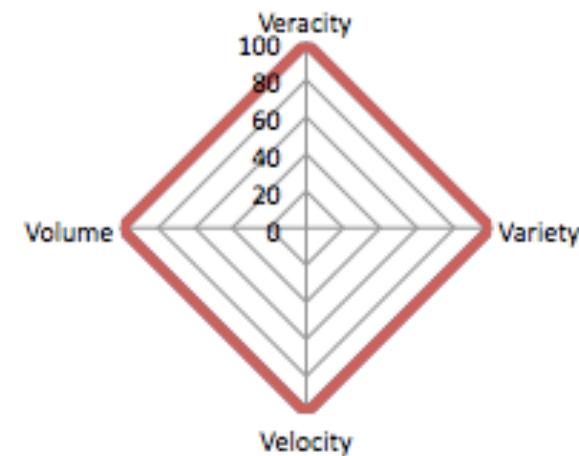
De gros volumes,  
Très variés,  
Vitesse rapide,  
Pas toujours juste  
Articles de presse ?



De gros volumes  
Très variés  
Blogs, Forums ?



De gros volumes,  
Disponibles rapidement  
Avec quelques erreurs  
Réseaux de capteurs ?



Google, Amazon, Twitter, etc

# Conclusion

---

- Rechercher les données disponibles pour aider à la prise de décision !
- Les outils sont encore à inventer !
- L'éducation est importante (Data Scientist)
- De très nombreux challenges



"Before I write my name on the board, I'll need to know how you're planning to use that data."