# Données massives : un renouvellement des problématiques sur la donnée, la mesure et l'hypothèse

**Bernard FALLERY**

MRM, Montpellier Recherche Management

bernard.fallery@univ-montp2.fr

AIM Lyon 2013

# Introduction



Vous saviez qu'Amazon prévoit les livres que vous allez aimer, mais :

- saviez-vous que si vous participez au réseau social Loopt, celui-ci peut prédire à 90% vos déplacements de demain ? (Ciarelli 2010)

- saviez-vous que VISA peut prévoir votre divorce en fonction des retards sur le paiement de vos achats à crédit ? (Ayres 2007)

# Introduction



La chaine Canadian Tire peut aussi différencier

- les bons payeurs (qui ont acheté des détecteurs de CO2, des graines de première qualité pour leurs oiseaux et … des coussinets de feutre pour leurs pieds de chaise)

- des mauvais payeurs (qui ont acheté de l'huile de moteur en promotion et … fréquentent le bar de la piscine de Montréal).

# Introduction



- eBay capture minute par minute les composants de chaque transaction pour repérer des patterns de comportements sur son site, où se vend par seconde pour plus de 2000 $ de produits

- Sur Amazon des logiciels comme AppEagle, SellerEngine... permettent de faire de la “tarification algorithmique” en fonction de votre profil (age, fidélité, équipement...) et de la localisation IP (distance d'un magasin physique, revenu de la zone...)

- Le gouvernement péruvien a lancé en 2012 un programme “big data” pour détecter la fraude fiscale

# Introduction

- Big data, de grands **programmes scientifiques** : en génomique (les cellules cancéreuses mutent différemment pour chaque individu), en physique (CERN, climatologie), en écologie...

- Big data, des **grandes entreprises** (IBM, Amazon-AWS, Google-BigQuery, SAP-Hana..), des **entreprises spécialisées** (Teradata, Jaspersoft, Pentaho...) et l'**Open Source** (Apache, Infobright, Talend...)

- Big data, des **Start-up** :
  Bionatics : dessins en 3D pour l'architecture
  Hariba Médical : simulation des produits en HD
  SafetyLine : gestion des risques du transport aérien
  KwypeSoft : mémoire des projets professionnels
  Vigicolis : gestion des livraisons e-commerce

# Plan



1. Big data, une rupture technologique dans l'étape de **gestion des données**

*Le débat : des données peuvent-elles être « brutes » ?*

2. Big data, une rupture technologique dans l'étape **d'analyse des données**

*Le débat : ne privilégier que le quantitativement mesurable?*

3. Big data, une rupture technologique dans l'étape **d'interprétation**

*Le débat : les chiffres parlent-ils « d'eux-mêmes » ?*

# 1. Une première rupture dans l'étape de gestion des données

- Démultiplication des **outils de collecte :** sur les individus et sur les objets (Web, RSN... Mobiles, capteurs...) **(1.1)**

- Démultiplication des **modèles de représentation :** bases NoSQL, programmation Map-Reduce (traitement parallèle) sous Hadoop (répartition sur différentes grappes de serveurs) **(1.2)**

- Démultiplication des **modèles de stockage** (Cloud Computing) : Big Query analyse 500 Go de données en 5 heures pour 200 euros **(1.3)**

## 1.1 Big data : une multiplication des outils de collecte

( 80 % des données sont aux Etats-Unis)

twitter

Twitter process ***7 TBs*** of data every day

: Logs & transactions

Facebook processes ***10 TBs*** of data every day

World Data Centre for Climate keeps ***220 TBS*** of Web data and ***9 PBs*** of auxiliary supporting data

Capital market data volumes grew ***1,750%,*** 2003-06

***30 billion*** RFID tags today (1.3B in 2005

***4.6 billion*** camera phones world wide

***900 million GPS*** devices sold annually by 2013

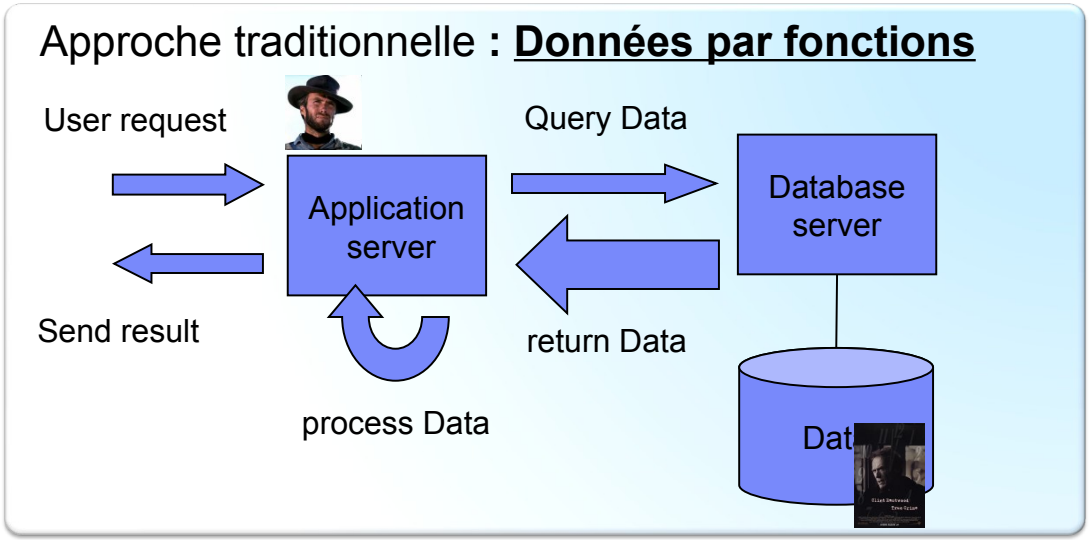
***76 million*** smart meters in 2009… 200M by 2014

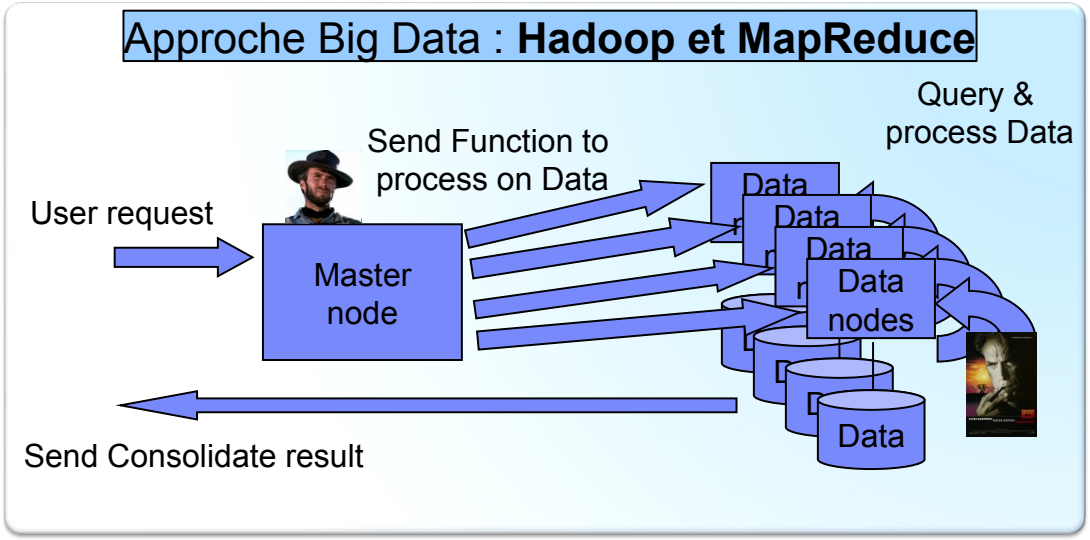

***2 billion*** people on the Web by 2011

Text, Blog, Weblog

# 1.2 Big data : une multiplication des modèles de représentation de données

## Exemple : Combien d'heures Eastwood apparait-il dans tous ces films ?




**Approche traditionnelle : tous les films transitent par le réseau**

- Les bases de données relationnelles sont sur des serveurs multiples

- les programmes d'analyse sont sur d'autres serveurs multiples

**Approche Big Data : seuls les programmes et la photo transitent par le réseau**

- Hadoop : les BD NoSQL (orientées colonnes, graphes, documents...) distribuent les données : 1000 noeuds par “scaling linéaire”

- MapReduce :les programmes s'executent en parallèle sur les noeuds où se trouvent les lots données (Map), puis les résultats sont agrégés (Reduce)

## 1.3 Multiplication des modèles de stockage

- **Le Cloud Computing :** accès via le réseau, à la demande et en libre-service, à des ressources informatiques partagées configurables (logiciels, données, plateformes, infrastructures).

**- Les supercalculateurs hybrides : L'exemple de HPC-LR**

Un cluster de calcul d'une puissance de 20,57 Teraflop :

- 84 **noeuds de calcul** bi-processeurs hexacore Intel
- 2 **noeuds large mémoire** Intel (80 coeurs/noeuds, 1 To RAM/noeuds)
- 6 **noeuds CPU/GPU** (bi-processeurs quadcore + 2 cartes NVIDIA)
- 4 **noeuds bi-processeurs** Cell
- 1 **noeud Power7** (16 coeurs)

- Baie de stockage de plus de 150 To utile
- Réseau Infiniband QDR
- General Parallel File System **GPFS (IBM**) : adressage de données réparties sur de nombreux supports physiques *(record IBM : 10 milliards de fichiers sur un système de stockage en 43 minutes).*

# Un débat renouvelé sur « la donnée »

Un nouveau discours : *« La masse de données peut compenser la qualité »*

**Les données peuvent-elles être brutes ?**

**Q'est-ce qu'une donnée ?**

- Accès à la réalité : sensation, perception, représentation
- Différences entre signal, donnée, information, connaissance
- Les données ne préexistent pas aux théories

**Les (big) data ne sont-elles pas toujours « cuisinées » ?**

- Les données sont toujours nettoyées par filtrage subjectif (question, partis-pris, extraction, variables, attributs...)
- Un jeu de données a toujours une certaine « qualité » / « faiblesse »
- Une donnée accessible n'est pas toujours une donnée publique
- Une donnée a toujours un contexte historique

# 2. Une deuxième rupture dans l'étape d'analyse des données

***Business Intelligence ou « Fouille de données »***
*- exploitent les données structurées d'une entreprise (requêtes OLAP, détection sans a priori des combinaisons de critères les plus discriminantes)*
*- avec les méthodes de classification (patterns) et de prédiction (modèles de scoring, ex : k-plus proches voisins)*

***Big Analytics* ou *« Broyage de données »***
- analysent des **données quantitatives complexes**, en temps réel, internes et externes, privées et publiques, des signaux faibles... **(2.1)**
- avec de nouvelles méthodes de **calcul distribué** : algorithmes inductifs, recherche de régularité, recherche de singularité, Web sémantique... **(2.2)**

# 2.1 L'analyse de données quantitatives complexes

**Big Data : une nouvelle opportunité de croiser les données**

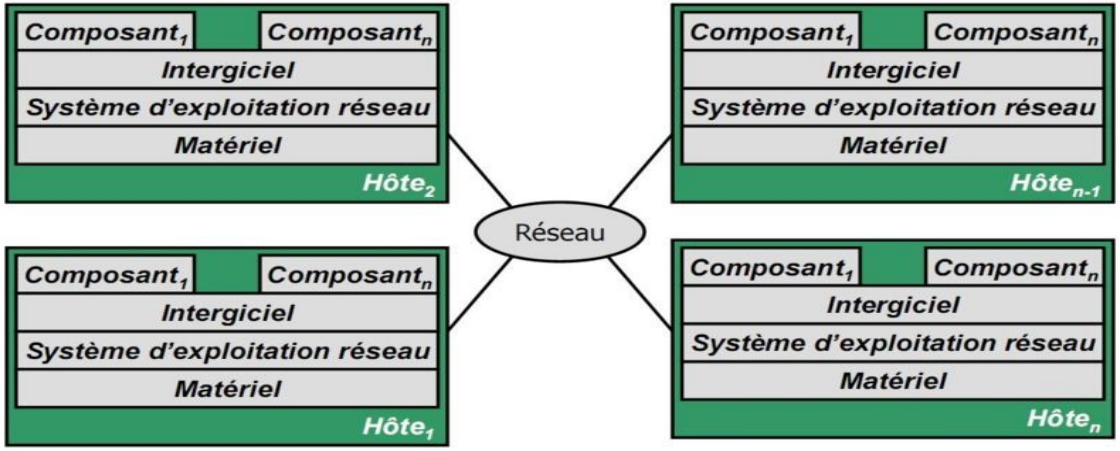
**Données complexes :**
Web : (SI entreprise x navigation Web) : *Web Mining*
Textes : (ADT Lexicale ou linguistique, TAL Langues )*Text Mining*
Images : (reconnaissance de forme, biométrie...) *Image Mining*

Données publiques ouvertes : *Open Data, Web des données*
Données géo-démographiques par îlot (*par adresses I.P.*)
Données sur les consommateurs (*Profils 360°*)

**Analyses complexes : les 4 « V »**
• Volume : stockages distribués et traitements parallèles
• Variété : intégration de données hétérogènes
• Vélocité : vitesse de capture et de traitement (stream computing)
• Variabilité : historisation des données (conservation, capture à chaque évolution...)

# 2.2 L'extension du calcul distribué

**Big data : du « Grid Computing » aux grappes de serveurs**




- Le « **Grid Computing »** sur des matériels hétérogènes : SETI sur les messages extra-terrestres, les défis sur les clés RSA, Décrypthon, ClimatePrediction...

- **Le calcul massif** sur des grappes de serveurs (Clusters) : Exploration de données, réduction de variables, régressions, réseaux neuronaux, forêts d'arbres de décisions, sévérité d'événements aléatoires, analyse inductive des situations-contextes ...

# Un débat renouvelé sur « la mesure »

Introduction
Donnée
Mesure
Hypothèse
Discussion et Conclusion

Un nouveau discours : *« les données massives offrent aux sciences humaines une opportunité : revendiquer le statut de **science quantitative aux méthodes objectives** »*

**Ne privilégier que le quantitativement mesurable?**

**Un nouveau débat « Quanti-Quali »** (Durkheim 1895 ...)
- Big Data, mais pas Whole Data : ni aléatoire (ex : contacts ou relations?), ni représentatif (ex : compte ou utilisateur Twiter?), ni complet ( ex : fausse alarme et non détection), ni complexe (feed-back, période ?)
- Le contexte demeure crucial, la carte n'est pas le territoire

**Un nouveau débat « objectif-subjectif »**
- Les outils donnent forme à la réalité qu'ils mesurent (Qui parle?)
- Une nouvelle fracture numérique (Big Data : accès, coûts ?)
- Peut-on tout mesurer sans consentement ? (éthique de la recherche)

# 3. Une troisième rupture dans l'étape d'interprétation des données

Un nouveau discours : *“Corrélation n'est pas causalité, mais cela n'a plus d'importance... laisser les algorithmes trouver les modèles que la science n'arrivait pas à trouver...” C. Anderson, 2008*

**La recherche sans modèles ? (3.1)**
- Les théories seraient définitivement incomplètes (cosmologie, médecine, traduction... comportements humains)
- Les variables causales seraient infinies : regarder d'abord les données mathématiquement et établir leur contexte ensuite ?
→ ***Observer, ~~théoriser~~ calculer, prédire l'observation suivante***

**De l'océan de données … à l'océan de corrélations ? (3.2)**
Peut-on construire quelque chose qui fonctionne, mais que nous ne comprenons pas?
Tout calcul donne des réponses, mais quelles sont les questions?

# 3.1 Exemples de recherches sans modèles



**L'équation d'Ashenfelter (2008) :**

Qualité du vin = 12,145 + 0,00117 Précipitations d'hiver + 0,0614 Moyenne des augmentations de température – 0,00386 Précipitations pendant la récolte

**Le séquençage des gènes de l'océan par G. Venter :**

Des séquences d'ADN différentes des autres dans la base (alertes statistiques) ont permis de découvrir des centaines de nouvelles espèces de bactéries, dont on ne sait encore rien.

# 3.2 De l'océan de données... à l'océan de corrélations



**Deux exemples :**

- *New England Journal of Medicine (2012) : corrélation très significative par pays entre la* ***consommation de chocolat*** *et le* ***nombre de prix Nobel***
- *En 2004 dans les* ***"swing states" (****Bush/Kerry), le scoring des électeurs a permis d'appeler par téléphone …. ceux qui possédaient un chat et une voiture à deux portes.*

**Deux questions :**

- De l'océan de corrélations … à **un océan de théories ?** (et de faux savoirs)
- Particulièrement dangereux en **sciences appliquées** ? (où l'essentiel n'est-il pas que "cela marche")

# Un débat renouvelé sur « l'hypothèse »

Un nouveau discours : *“Nous pouvons désormais analyser les données sans faire d'hypothèses sur ce qu'elles vont produire» C. Anderson, 2008*

**Les chiffres peuvent-ils parler d'eux-mêmes?**

**Le débat sur la méthode expérimentale**

L'hypothèse permet d'aller au-delà de l'extrapolation

L'hypothèse permet de reconnaitre sa subjectivité

L'hypothèse permet l'enrôlement (Sociologie de la traduction)

Les données ne peuvent que réfuter une hypothèse, pas la valider

**Le débat « prédire ou expliquer »?**

- DES explications (falsification) et DES prédictions (logiques et dynamiques, téléologiques) : Offre/Demande, l'Evolution...
- L'ingénierie par pure induction ? (Marketing, Criminologie, Finance)

**Conclusions sur la rupture technologique « Big data » :**

- elle est probablement très forte à l'étape de gestion de données
- elle n'est qu'une évolution au niveau de l'analyse de données
- elle est probablement dangereuse à l'étape de l'interprétation

*(ces trois ruptures concernent aussi les petites organisations)*

**Conclusions sur le renouvellement des problématiques :**

- le débat sur « la donnée » sera plus difficile pour les constructivistes
- le débat sur « la mesure » sera plus difficile pour les SHS
- le débat sur « l'hypothèse » sera plus difficile pour les sciences appliquées (Médecine, Gestion..)

*(la pluridisciplinarité, obligatoire, sera probablement difficile)*

**A. Einstein :** *C'est la théorie qui décide de ce que nous sommes en mesure d'observer*

**Big Data :** *Est-ce la théorie qui décide de ce que nous sommes en mesure d'observer ?*

**Ayres I. (2007)**
Super crunchers : Why Thinking-by-Numbers Is the New Way to be Smart, Bantam Books, N.Y 2007

**Anderson, C. (2008)**
The End of Theory, Will the Data Deluge Makes the Scientific Method Obsolete?
*Débat : Edge, 25 July 2011 www.edge.org/discourse/the_end_of_theory.html*

**Boyd D., Crawford K. (2011)**
Six Provocations for Big Data,  Symposium on the Dynamics of the Internet and Society, September 2011
*Traduction : www.internetactu.net/2011/09/23/big-data-la-necessite-d%E2%80%99un-debat/*