



Une approche interdisciplinaire des grandes masses de données (Défi Mastodons)

www.cnrs.fr

Mokrane Bouzeghoub
DAS INS2I / MI

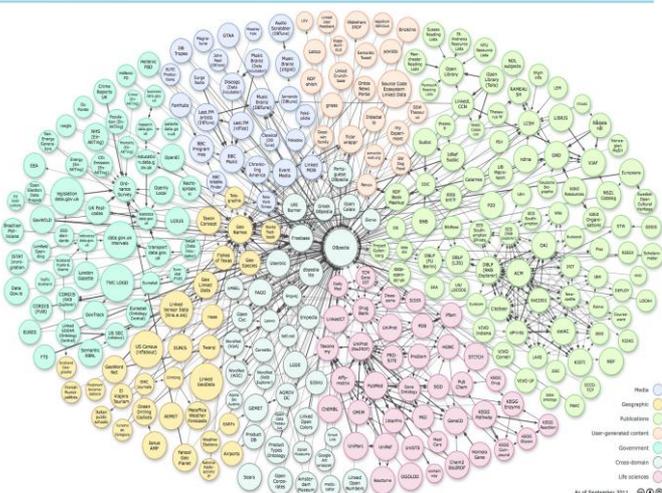
Ecole de L'Innovation Thérapeutique Ariis / Aviesan – 13 Juin 2014



Emergence du Big Data Exemple : Linked Open Data

**Accès à
plusieurs BD
scientifiques et
culturelles
interconnectées
sur le Web**

**Ini§ée en 2007 avec
une dizaine de
sources de données
interconnectées**



As of September 2011

Aujourd'hui, plusieurs centaines de sources connectées et ouvertes



cnrs
dépasser les frontières

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



Grandes Conf du domaine

VLDB

Big Data

XLDB

Very Big Data

Massive Data

Data Deluge

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 <i>The Economist</i>
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

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.
Source: *The Economist*

CNRS l'interdisciplinarité

Mokrane Bouzeghoub

3



cnrs
dépasser les frontières

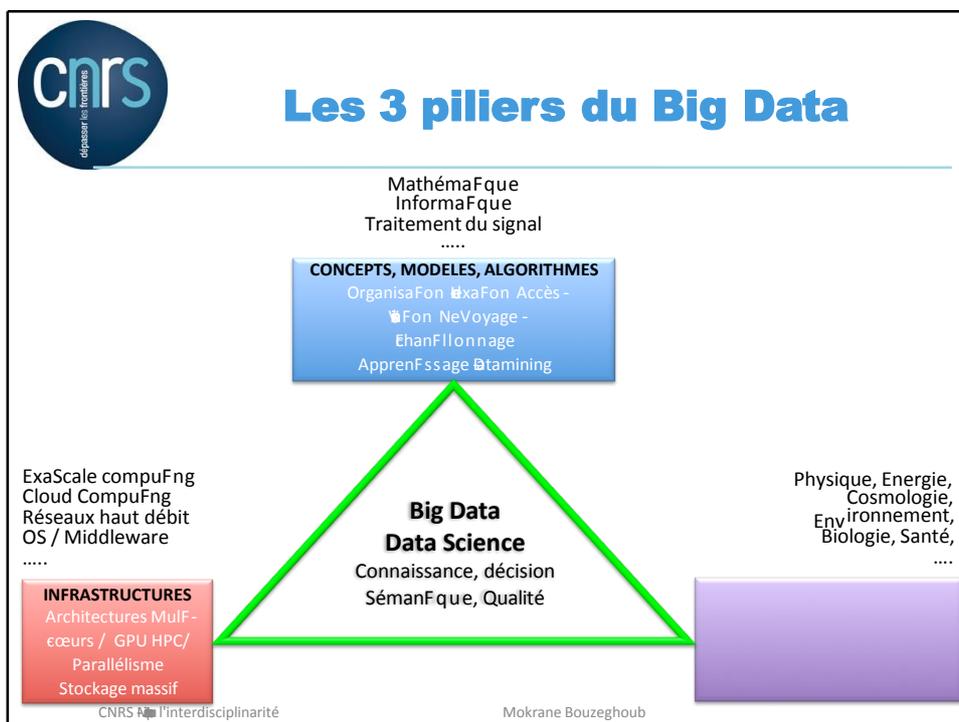
Les grandes questions du Big Data

- **La science est-elle dans les masses de données ?**
 - La valeur de ces données réside dans les indicateurs, les paramètres et les règles/lois qui peuvent en être dérivés (**connaissance**)
 - Ces données sont importantes non seulement en raison de leur quantité mais aussi en raison des relations existantes entre elles (**sémantique**)
 - Les données peuvent être source de **plus de science** mais aussi source de bruit et de pollution (**qualité, hétérogénéité, manipulation**)
- **Les masses de données nous parlent-elles de notre société ?**
 - Nous disent-elles quelque chose que nous ne sachions déjà ?
 - Diront-elles quelque chose de nous aux générations futures ?
 - Ont-elles une objectivité en elles-mêmes ou sont-elles biaisées par des transformations subjectives ?
- **Les masses de données génèrent-elles une valeur économique ?**
 - Quels sont les secteurs privilégiés ?
 - Quel retour sur investissement ?
 - Quel rôle pour ces données (matière première, produits dérivés, capital, ...) ?
 - Quel statut pour ces données (propriété privée, domaine public, objet commercial) ?

CNRS l'interdisciplinarité

Mokrane Bouzeghoub

4



cnrs dépasser les frontières

La complexité multidimensionnelle du Big Data

- **La Volumétrie**
 - Un défi pour les architectures de stockage (au delà du PB)
- **La Variété**
 - Diversité des contenus
 - Forte hétérogénéité des formats et des données
- **La Vitesse**
 - Défi pour les nouveaux réseaux de communication
 - Nouveaux modèles de calcul sur des données en flux
- **La Validité / Vérité**
 - Qualité des sources de données: fraîcheur, exactitude, ...
 - Qualité des processus de production/transformation

WWW

CNRS et l'interdisciplinarité Mokrane Bouzeghoub 6



Les grands challenges scientifiques du Big Data

- **Stockage dans le Cloud**
 - Performance des accès, disponibilité
 - Sécurité des données et des traitements
- **Complexité du calcul**
 - Analyse en temps réel de flux continus de données émanant de différentes sources
 - Requêtes multidimensionnelles sur des grands ensembles de données
- **Sémantique des données**
 - Indexation sémantique (ontologies), indexation participative (folksonomies)
 - Extraction et interprétation de connaissances
- **Consommation d'énergie**
 - Ressources à énergie limitée (ex. capteurs)
 - Optimisation du transfert des données
- **Impact sociétal**
 - Protection de la vie privée, Droit à l'oubli
 - A qui appartiennent les données, les connaissances?

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



Caractéristiques du domaine

- **Un domaine très vaste, !**
 - en interaction permanente avec les autres disciplines scientifiques!
- **Un domaine qui se repositionne périodiquement!**
 - En revisitant ses solutions à la lumière de nouvelles technos et de nouvelles idées!
 - En intégrant de nouveaux besoins et de nouveaux problèmes!
- **Une recherche dominée (ou presque) par des labos industriels : !**
 - Google, Facebook, Yahoo!, Amazon, IBM, Oracle, Microsoft ...!



Quelques initiatives en Big Data

- **USA : Plusieurs acteurs dont**
 - Gouv't US: Big Data Research and Development Initiative (Mars 2012)
 - ✓ 250M\$ / an dont 60 pour les projets de recherche
 - ✓ mis en œuvre par NSF, NIH, DOD, DOE, USGS)
 - Accel Partners: fond d'investissement 60 M\$ / an de soutien à la création de startups dans le Big Data
- **UK: Plusieurs initiatives dont**
 - ESRC Big Data Network (2012) : 3 phases, PHASE 2 AVR 2013: 60M£.
 - BBSRC (2012): 75 M£ pour améliorer la disponibilité des Big Data
- **France**
 - PIA: Appel 'Cloud Comp & Big Data Ministère de l'Industrie (juillet 2012): 25 M€
 - **CNRS: Initiative interdisciplinaire (Mastodons): 700K€/an sur 4/5 ans?**

CNRS l'interdisciplinarité

Mokrane Bouzeghoub

9



Objectifs du défi Mastodons

Produire des concepts et des solutions
qui n'auraient pu être obtenus
sans coopération entre les différentes disciplines



Favoriser l'émergence
d'une communauté scientifique interdisciplinaire
autour de **la science des données**,
et produire des solutions originales sur
le **périmètre des données scientifiques**.

CNRS l'interdisciplinarité

Mokrane Bouzeghoub

10



Focus de l'appel Mastodons

- **Stockage et gestion de données (par exemple, dans le Cloud), sécurité, confidentialité!**
- **Calcul intensif sur des grands volumes de données parallélisme dirigé par les données!**
- **Recherche, exploration et visualisation de grandes masses de données!**
- **Extraction de connaissances, datamining et apprentissage!**
- **Qualité des données, confidentialité et sécurité des données!**
- **Problèmes de propriété, de droit d'usage, droit à l'oubli!**
- **Préservation/archivage des données pour les générations futures!**



Les critères de sélection

- **Vision scien\$ifique de l'équipe/consor\$um sur les thèmes du défi**
- **Les verrous scien\$ifiques et les axes de recherche à moyen terme, avec un focus par\$culier sur la première année**
- **Les acquis scien\$ifiques dans le domaine ou dans un domaine connexe suscep\$ble de contribuer aux problèmes scien\$ifiques ou sociétaux posés (publica\$ons significa\$ves, projets passés ou en cours, applica\$ons réalisées, logiciels, brevets...)**
- **Les différentes disciplines impliquées et leurs contribu\$ons respec\$ves au projet**
- **Une liste de 3 à 5 chercheurs seniors impliqués de façon significa\$ve dans la recherche.**

 ***l'interdisciplinarité doit être une réalité et pas un alibi***



Indicateurs de suivi

- Pérennité de la coopération
- Publications communes
- Coencadrement de thèses
- Plateformes de test et d'expérimentation
- Montage et soumission de nouveaux projets
- Dynamique pour faire émerger une communauté interdisciplinaire sur la science des données.



Mastodons : Chiffres clés

- Défi lancé en 2012, avec un second appel en 2013
- Projets de 3 à 5 ans
- Budget : environ 700 à 850 K€/an
- Nb de soumissions: 57
 - Nb d'UMR impliquées: + 100, Couvrant les 10 instituts
- Nb de projets retenus: 20
 - Nb d'UMR impliquées: 69, couvrant les 10 instituts
 - Nb de CH/EC impliqués: près de 300
 - Montant alloué/projet : 30 à 80 K€
- Partenaires hors CNRS
 - INRIA, INRA, IRSTEA, INSERM, CEA, ONERA
 - Universités et écoles



Types de données visés dans les projets retenus

- **Cosmologie, astrophysique**
 - Dynamique de la Cartographie céleste
- **Sciences de la terre et de l'univers (traitement d'images)**
 - Modélisation, déformation de la croûte terrestre
- **Environnement, climat, biodiversité**
 - Simulation, intégration, fusion de données
- **Biologie**
 - Génome, séquençage, phénotypage
- **Réseaux sociaux**
 - RI, analyse d'opinions, santé

CNRS  l'interdisciplinarité

Mokrane Bouzeghoub

15



Deux ans après...

Gros projets phares

- **PetaSky+Gaia +Amadeus**
 - Cosmologie
- **Aresos**
 - Réseaux sociaux
- **Phénotypage, Sabiod**
 - Biologie végétale, Écologique

Projets ciblés excellents

- **Comotex**
 - Cde Tps réel de syst op\$que
- **Display**
 - Distr proc. For VLA in Radioastronomy
- **MesureHD**
 - Mesures hautes résolution
- **Prospectom**
 - Etude interac\$ve des protéomes par appren\$ssage stat. et intégr de données spectrométriques

+ Un projet émergent sur le crowdsourcing: CrowdHealth

CNRS  l'interdisciplinarité

Mokrane Bouzeghoub

16



Mastodons : La suite ...

- **Comment pérenniser la communauté**
 - **Réflexion générale sur les regroupements de projets**
 - ✓✓Thématique
 - ✓✓Par domaine d'application
 - **Structuration et animation de la communauté 'Big Data'**
 - ✓✓Emergence d'un GDR « Big Data, Science des données »
- **Comment la financer au delà du programme CNRS**
 - **CNRS, au delà de 2015?**
 - **ANR ?**
 - **COST / H2020 ?**
 - **Autre initiative ?**

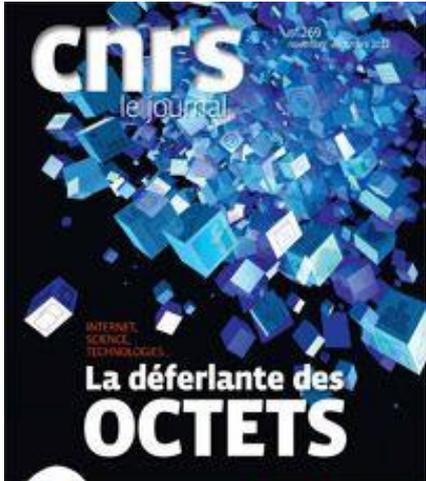


Conclusion

- **La recherche en Big Data ne peut être fructueuse sans un rapprochement des chercheurs des grands centres de production et d'exploitation des données (existants ou à créer)**
 - **Avec un soutien fort en ingénierie**
 - **Une véritable interdisciplinarité**
 - **Un code clair sur l'accès aux données et leur utilisation**

cnrs dépasser les frontières

Publications CNRS récentes



La déferlante des OCTETS
INTERNET, SCIENCE, TECHNOLOGIES

The BIG DATA REVOLUTION
international magazine

19

CNRS-Mission pour l'interdisciplinarité

Mokrane Bouzgehoub

cnrs dépasser les frontières

Journal du CNRS



CNRS LE JOURNAL

VIVANT MATIÈRE SOCIÉTÉS UNIVERS TERRE NUMÉRIQUE

MES TRÈMES

20

CNRS Mission pour l'interdisciplinarité

Mokrane Bouzgehoub