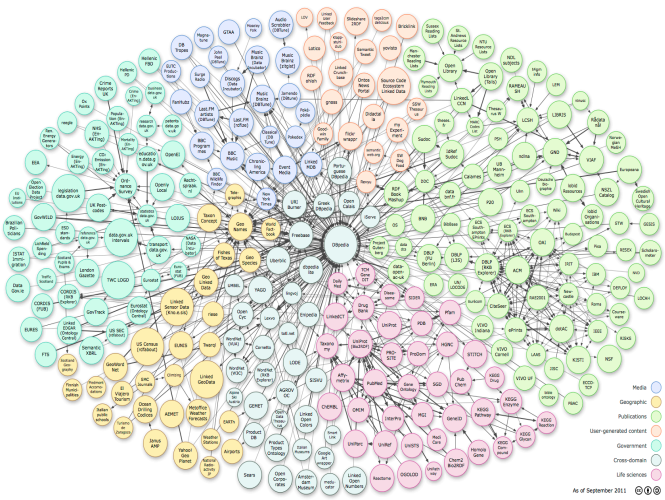
Une approche interdisciplinaire des grandes masses de données (Défi Mastodons)

www.cnrs.fr

Mokrane Bouzeghoub
DAS INS2I / MI

Emergence du Big Data
Exemple : Linked Open Data

Accès à plusieurs BD scientifiques et culturelles interconnectées sur le Web



Initiée en 2007 avec une dizaine de sources de données interconnectées

Aujourd'hui, plusieurs centaines de sources connectées et ouvertes

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

VLDB

XLDB

Big Data

Very Big Data

Data Deluge

Massive Data

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, …

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 patterns 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-value scientifique 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 publique, objet commercial)?

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 Vélocité**
  - Défi pour les nouveaux réseaux de communication
  - Nouveaux modèles de calcul sur des données en flux
- **La Validité / Véracité**
  - Qualité des sources de données: fraîcheur, exactitude, …
  - Qualité des processus de production/transformation

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/Tera-octet 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!, Amazone, IBM, Oracle, Microsoft …

Quelques initiatives en Big Data

- USA : Plusieurs acteurs dont
  - Gouvt 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?

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.

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 scientifique de l'équipe/consortium sur les thèmes du défi
- Les verrous scientifiques et les axes de recherche à moyen terme, avec un focus particulier sur la première année
- Les acquis scientifiques dans le domaine ou dans un domaine connexe susceptible de contribuer aux problèmes scientifiques ou sociétaux posés (publications significatives, projets passés ou en cours, applications réalisées, logiciels, brevets...)
- Les différentes disciplines impliquées et leurs contributions respectives au projet
- Une liste de 3 à 5 chercheurs seniors impliqués de façon significative dans la recherche.

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

Indicateurs de suivi

- Pérennité de la coopération
- Publications communes
- Co-encadrement 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 croute terrestre
- Environnement, climat, biodiversité
  - simulation
- Biologie
  - Génome, phénotypage
- Réseaux sociaux
  - RI, analyse d'opinions, santé

Deux ans après…

Gros projets phares

- PetaSky+Gaia +Amadeus
  - Cosmologie
- Aresos
  - Réseaux sociaux
- Phénotypage, Sabiod
  - Biologie végétale, Bio-acoustique

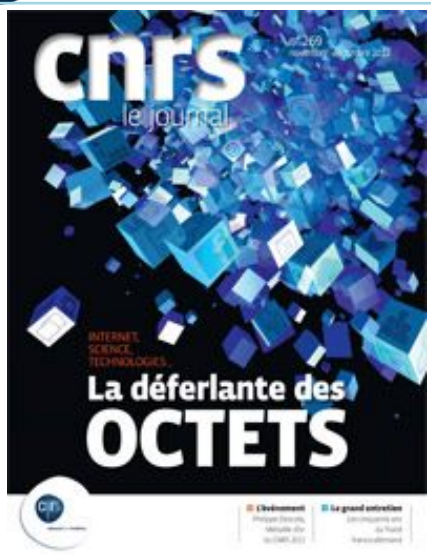
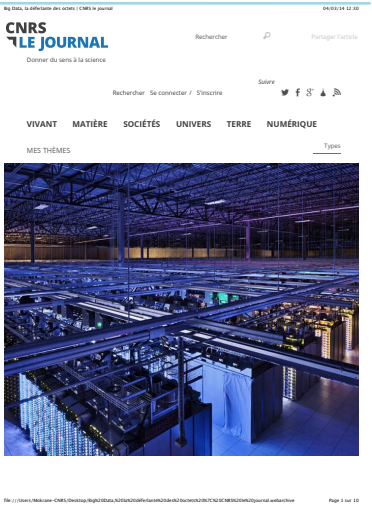
Projets ciblés excellents

- Comotex
  - Cde Tps réel de syst optique
- Display
  - Distr proc. For VLA in Radioastronomy
- Mesure-HD
  - Mesures hautes résolution
- Prospectom
  - Etude interactive des protéomes par apprentissage stat. et intégr de données spectrométriques

+ Un projet émergent sur le crowdsourcing: CrowdHealth

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