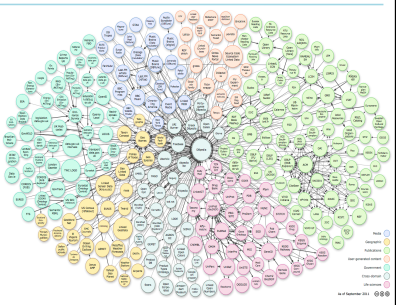
# Le Défi Mastodons
# Une approche interdisciplinaire des grandes masses de données

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**

## 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)?

## 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

## 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**
  - Govt 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): 800K€/an sur 4/5 ans?

## Le Défi Mastodons : Objectifs

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, indexation et gestion de données (par exemple, dans le Cloud),
- 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.

## Mastodons : Chiffres clés

- Défi lancé en 2012, avec un second appel en 2013
- Projets de 3 à 5 ans avec un budget de 700 à 885 K€/an
- Nb de soumissions: 57
  - Nb d'UMR impliquées: + 100, Couvrant les 10 instituts
- Nb de projets retenus: 20 +1
  - Reste 17 projets en janvier 2014, cible janvier 2015: 10 projets
- Degré de pénétration dans les labos
  - Nb d'UMR impliquées: 69, couvrant les 10 instituts
  - Nb de CH/EC impliqués: près de 300
- Montant alloué/projet
  - 30 à 120 K€ (projets ayant fusionné)
- Partenaires hors CNRS
  - INRIA, INRA, IRSTEA, INSERM, CEA, ONERA, Universités et écoles

## Les projets retenus (par grands domaines)

- Physique des particules et astrophysique: 2
- Sciences de la terre et de l'univers: 5
- Environnement, climat, biodiversité: 4
- Biologie: 3
- Réseaux sociaux: 2
- Préservation des données: 1
- Traitement d'images : 2
- Apprentissage statistique : 1
- Qualité des données : 1

## Trois ans après…

**Gros projets phares**

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

**Projets ciblés excellents**

- Comotex
  - Cde Tps réel de syst optique
- CrEDIBLE
  - Int. sémantique de données médicales
- Display
  - Distr proc. For VLA in Radioastronomy
- Mesure-HD
  - Mesures hautes résolution
- Prospectom
  - Apprentissage stat. et intégr de données spectrométriques

**2 projets sur le crowdsourcing:**
CROWD-BIODIV: Evolution de la biodiversité
CROWD-HEALTH: Corrélation données nutritionnistes et maladies

## Projet Aresos : Analyse de grands réseaux socio-sémantiques

CAMS - INSMI, EHESS, Paris
CSI - INSHS, Ecole des Mines, Paris
IRIT - INS2I, U. Toulouse 3
LATTICE - INSHS, ENS/ U. Paris 3
LIG - INS2I, UJF, Grenoble
LIP6 - INS2I, UPMC, Paris
IRISA, INS2I, U. Rennes 1
GIS ISC-PIF, INSHS

- Objectifs : qui parle, de quoi, comment
  - Reconnaissance d'acteurs
  - Analyse sociologique
- Recherche d'information dans les microblogs
  - Identification de thématiques
- Recommandation collaborative
  - CrowdIndexing, tagging social

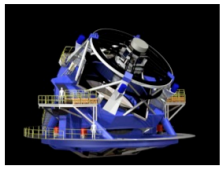
Défi MASTODONS - Projet ARESOS 15

## Projet SeqPhénoHD : Séquençage & Phénotypage Haut Débit

- Info et bio-info
  **LIRMM, LIFL, IRISA**
- Phénotypage
  **INRA**
- Génome
  **France Génomique**
- Biologie-environnement
  **ISEM**

- Etude du comportement des plantes, de différents génomes,
  - Densité végétation (nb de feuilles)
  - Croissance (rapidité, hauteur, encombrement, …)
- selon les évolutions de leur environnement
  - Température,
  - Humidité,
  - Lumière/Ensoleillement
- Exemple
  - 400 génomes
  - 3 à 10 plants par génome
  - $10^5$ informations / jour

## Projet PetaSky : observation astronomique grand champ (LSST)



LIMOS (Clermont-Fd)=> **F. Toumani**
LIRIS (Lyon)
LPC (Clermont-Fd)
APC (Paris)
LAL (Paris)
Centre de Calcul de l'IN2P3/CNRS

15 CEC, 8 ITA, 2 Doct.

- Gestion des données scientifiques dans le domaine de la cosmologie et l'astrophysique
- Des dizaines de milliers de milliards d'observations photométriques sur des dizaines de milliards d'objets
  - 3 Milliards de sources
  - 1-10 Millions d'évènements par nuit
  - 16 TB chaque 8 heures avec un taux de 540 MB/seconde

Estimation en fin de projet : 400 000 Milliards de tuples (différentes versions des données sans prise en compte de la réplication), =60 PB

## 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.

## Indicateurs de suivi : quelques chiffres (Janv 2014)

- Publications communes
  - 25 publications de haut niveau
  - 5 workshops internationaux organisées
  - Plusieurs journées d'études propres aux projets ou inter-projets
- Levier pour lancer d'autres projets
  - France: 5 ANR + 1 PIA Big Data
  - Europe: 1 Flagship + 2 FP7 soumis
  - Projet COST BIG-SKY-EARTH : Big Data Era in Sky and Earth Observation
    - ✓ 16 pays partenaires dont les membres de PetaSky + Gaia
- Autres impacts (projet Sabiod)
  - Bird Challenge: Identify bird species from continuous audio recordings
  - Expédition Goélette TARA: collecte et traitement de données sur la pollution en méditerrannée

## Perspectives 2015

- Poursuivre la structuration de la communauté
  - Via le financement des gros projets (regroupement, renforcement)
  - Via le nouveau GDR MaDICS (Christine Collet)
- Susciter de nouveaux projets sur
  - La sécurité et la protection de la vie privée
  - La visualisation des données
  - Le crowdsourcing (aspects SHS)
- Année thématique pour INS2I
  - Coloriage de postes de CR1/CR2
  - Soutien aux plateformes (ingénieurs)
  - Soutien aux recherches théoriques (AAP)

## 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
- Structuration de la communauté
  - Émergence de sites de références

## Big Data : un enjeu pour le CNRS




## Big Data = Big Topic





23