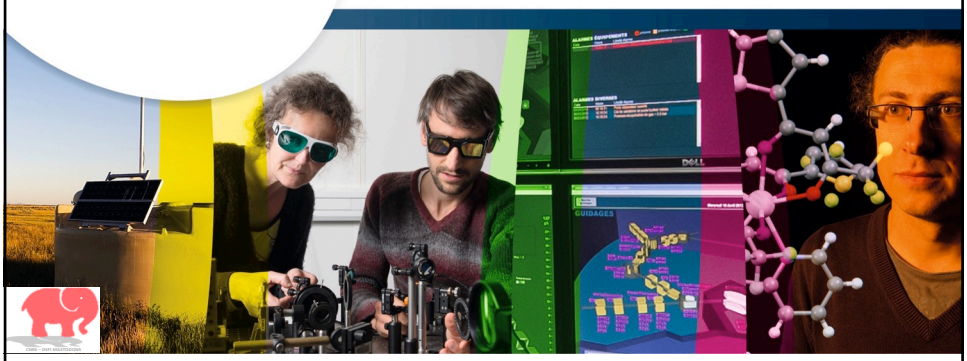
Big Data et Sciences des Données
- Le défi Mastodons –

Mokrane Bouzeghoub
CNRS / INS2I & MI

www.cnrs.fr



L'Humanité accumule des données depuis des millénaires

- **Ce qui change**
  - les technologies d'acquisition, de stockage et de traitement
  - La numérisation intensive de la société
- **Ce qui est nouveau**
  - l'accélération du phénomène
  - La diversité et la multiplicité des usages

Une accélération vertigineuse

- En 2 jours nous avons produit plus de données que l'Humanité n'en a produit en 2 millions d'années !

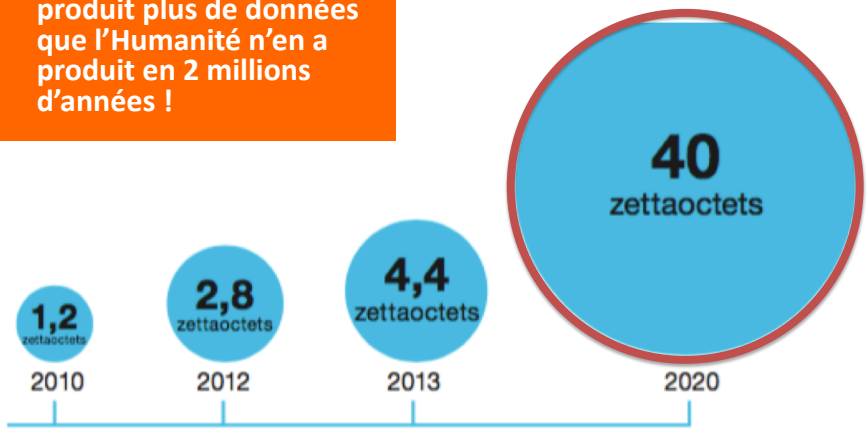


Fig. 1 – Croissance passée et prévisible de la quantité de données créées dans le monde. ■

How big is big ?

O	1	1
Kilo-O	1 000	10^3
Méga-O	1 000 000	10^6
Giga-O	1 000 000 000	10^9
Téra-O	1 000 000 000 000	10^{12}
Péta-O	1 000 000 000 000 000	10^{15}
Exa-O	1 000 000 000 000 000 000	10^{18}
Zetta-O	1 000 000 000 000 000 000 000	10^{21}
Yotta-O	1 000 000 000 000 000 000 000 000	10^{24}

Aucun domaine n'échappe à l'avalanche des données

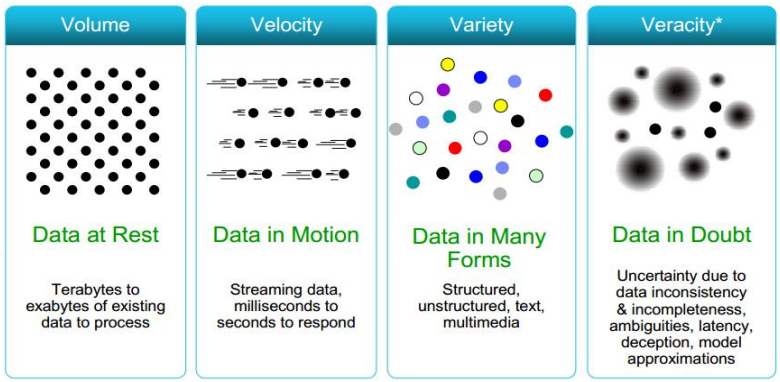
- **Commerce et les affaires**
  - SI d'entreprise, Banques, transactions commerciales, systèmes de réservation, …
- **Gouvernements et organisations**
  - Lois, réglementations, standards, infrastructures, ….
- **Loisirs**
  - Musique, vidéo, jeux, réseaux sociaux…
- **Sciences fondamentales**
  - Astronomie, physique et énergie, génome, …
- **Santé**
  - Dossier médical, sécurité sociale,…
- **Environnement**
  - Climat, dév durable, pollution, alimentation,…
- **Humanités et Sciences Sociales**
  - Numérisation du savoir (littérature, histoire, art, architecture), données archéologiques…

Une grande variété de données

- **Données de références**
  - **Cadastre, SIG, catalogues de produits, …**
- **Données d'observations**
  - **Satellites, capteurs, expériences scientifiques, …**
- **Données transactionnelles**
  - **Transactions commerciales, requêtes BD/Web, …**
- **Données sociales**
  - **Web, Facebook, Twitter, Crowdsourcing, …**
- **Données du patrimoine**
  - **Culture, architecture, publications, ...**
- **...**

Représentées sous forme
de tables, de graphes, de textes, d'images ou de flux vidéo

Autres caractéristiques



Nouvelles architectures de stockage et d'interopérabilité

Nouveaux modèles de calcul sur des flux (Défi pour les réseaux de communication)

Nouveaux outils de Nettoyage, transformation et Fusion de données

Nouveaux modèles de qualité (données & processus de traitement)

http://www.datasciencecentral.com/profiles/blogs/data-veracity

Big Data – Définitions différentiées

- **Chercheur informaticien**: c'est la frontière au delà de laquelle les outils dont je dispose ne permettent pas le traitement effectif des données, en raison de leur volume ou en raison de méthodes inadaptées

- **Non-informaticien :** C'est la mise en œuvre de nouvelles méthodes de traitement et d'analyse de données susceptibles de créer une rupture dans la façon de pratiquer le métier d'analyste ou de chercheur

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 …**

Les grands challenges du Big Data

- **Stockage et préservation des données**
  - Performance des accès, disponibilité des données
  - Protection des données, qualité 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, apprentissage profond
- **Impact social et économique**
  - Protection de la vie privée, Droit à l'oubli
  - Droits de propriétés, droits d'exploitation
  - **Economie d'énergie, écologie**
    - **coût du stockage, coût de transfert**

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

Le Big Data n'est pas gratuit

- Consommation électrique des data centers
  - Entre 600 TWh et 1 500 TWh, soit environ 10 % de l'électricité produite dans le monde, équivalent de la production de 180 réacteurs nucléaires
- Coût d'extraction des matières premières (terres rares)
  - non durable et souvent non éthique
- Coût de production et de renouvellement des équipements
  - Accélération de l'innovation
- Coût de recyclage des déchets
  - Un cycle partiellement maîtrisé

La recherche en big data est un continuum de la recherche sur les données

Plusieurs décennies de recherche et d'innovation

- Stockage, indexation, distribution, data cloud
- Requêtes continues, requêtes approximatives
- Requêtes avec préférences, requêtes skyline
- Analyse de flux, agrégation en ligne, algèbre OLAP
- Intégration de données, systèmes de médiation, ETL
- Fouille de données, découverte de motifs
- Exécution distribuée et parallèle de requêtes
- …

Des big data à la science des données

www.cnrs.fr

Perspective historique

- **Au cours de l'histoire, les plus grandes transformations se sont produites lorsqu'une ressource jusque-là rare est devenue abondante**
  - **L'émergence et le développement de l'agriculture (plus de nourriture)**
  - **l'invention de l'imprimerie (plus d'instruction, de culture)**
  - **la révolution industrielle (plus de produits, de mobilité)**
  - **Internet et le Web (plus de communication et plus de services)**
- **Ces « révolutions » ont systématiquement généré**
  - **de nouvelles structures sociales**
  - **de nouvelles formes de gouvernance**
  - **de nouvelles sources de richesse**
  - **de nouvelles opportunités**
  - **… mais aussi de nouvelles inégalités et de nouveaux risques.**

FR-UK TASK FORCE ON DATA-DRIVEN ECONOMY

Mokrane Bouzeghoub 14

Spécificité de la ressource « donnée »

On doit investir dans les données de la même manière que nous investissons dans les routes, les chemins de fer et d'autres infrastructures publiques.
(FR-UK Report on Data-Driven Economy)

- l'importance stratégique des données n'est pas encore reconnue par tous
  - Peuvent faire l'objet d'un usage partagé
  - Ne sont pas détruites après utilisation
  - La production de données est quasi infinie
- les données n'ont pas toujours de valeur intrinsèque
  - Leur valeur résulte des services, produits et connaissances qui génèrent de nouveaux modèles économiques et de nouvelles activités
  - leur utilisation répétée peut augmenter leur valeur et conduire à de nouvelles formes de ressources (métadonnées, connaissances)

De nouveaux types de traitements sur les données

- Sur les données du Web ou des réseaux sociaux (Facebook, Twitter, Instagram, ...).
  - Ces données ne résultent d'aucune production scientifique ni d'aucune enquête, ne correspondent à aucune observation, mais pourtant elles sont là et offrent, par leur volume et leur dynamique un champ d'investigation tout à fait inédit
- Emergence de nouvelles applications
  - Analyse d'opinions, identification de leaders, détection de signaux faibles, recommandations commerciales, …
  - À la portée du plus grand nombre
- … Et un usage réflexif par les algorithmes
  - amélioration des algorithmes d'apprentissage et de recherche d'information, en faisant un retour d'expérience à ces algorithmes.

www.cnrs.fr

Avec de nouveaux outils
Fouille de données, apprentissage, visualisation, sciences participatives (annotations,…)

Précurseur de la science des données

N'oublions pas que cette problématique a été mise en lumière par Jim Gray, en janvier 2007, disparu en mer deux semaines plus tard.

« The availability of very large amounts of data and the ability to efficiently process them is changing the way we do science. »

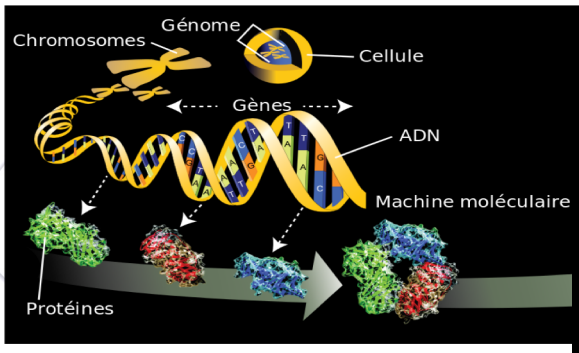
Science paradigms[1]

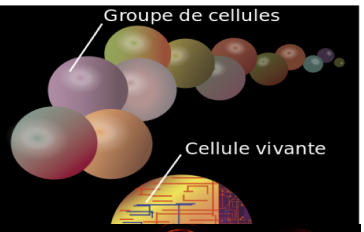
1. **Empirical description of natural phenomena**
2. **Theoretical science: models and generalization**
3. **Computational science: simulation of complex phenomena to validate theories**
4. **Data Intensive science : collecting and analyzing large data volume**

Turing Award

[1]Jim Gray, eScience Talk at NRC-CSTB meeting Mountain View CA, 11 January 2007.

Science des données et l'infiniment petit





- **Pour observer et comprendre le Vivant**
  - Génétique, Phylogénie, Biologie évolutive…
- **Pour comprendre la physique des particules**
  - LHC, énergie

Science des données et l'infiniment grand

LSST
- 1-10 Millions d'évènements observés par nuit
- 16 TB chaque 8 h

Pour répondre à des questions fondamentales
- Que s'est-il passé juste après le Big Bang
- Comment les galaxies sont formées
- Quelles lois physiques gouvernent l'Univers
- Comment évolue notre Univers

Science des données dans le quotidien à l'ère du Net

- **Internet a initié une nouvelle ère commerciale**
  - **En utilisant nos données personnelles et nos interactions (Les likes, les tag, les préférences, Les données personnelles)**
  - **En exploitant ces données pour de nouvelles opportunités commerciales**

Les réseaux sociaux deviennent un observatoire de la Société
- **Qui parle de quoi ?**
- **Quels sont les thèmes de discussion majeurs ?**
- **Quels sont les thèmes émergents ?**
- **Qui sont les leaders d'opinion ?**
- **Peut-on détecter des signaux faibles ?**
  - Suicide, attentat, pédophilie, ….

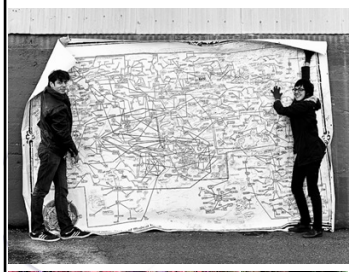
Science des données dans l'économie et le commerce

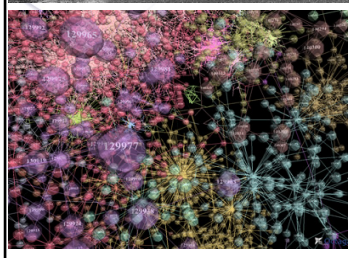
- En 2014, un rapport du McKinsey Global Institute a conclu, que le Big Data pourrait rapporter 3 000 Mds $ chaque année dans seulement 7 secteurs d'activité
- En 2015, la Commission européenne a estimé que la taille cumulée du marché direct des données publiques ouvertes devrait s'élever à 325 Mds € en Europe pour la période 2016-2020
- En 2015, 61 % des sociétés françaises estimaient que le Big Data était devenu l'un des principaux moteurs de croissance, aussi important pour elles que leurs produits et services existants
- en 2018, le marché français du Big Data devrait atteindre 652 M€, soit une hausse de 129 % par rapport à 2014, malgré une conjoncture défavorable

Quelles techniques pour faire parler les données ?

- Des outils d'analyse statistiques
- Des algorithmes de filtrage
- Une algèbre et des langages de manipulation
- Des algorithmes d'apprentissage (machine-learning, deep-learning)
- Des algorithmes de fouille de texte
- Des algorithmes d'analyse d'images
- Des algorithmes de traitement du signal (flux audio ou vidéo)
- …

Mais très souvent … seulement une exploration visuelle

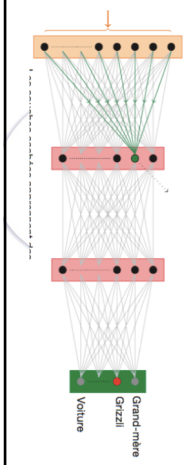




visualcomplexity.com/vc

- Pour observer des phénomènes non perceptibles à l'œil nu ou à l'aide d'instruments
  - durant la simulation par exemple
  - Interaction entre faisceaux de particules
- Pour analyser à un niveau macro des interactions entre objets
  - Phylogénie
  - Mouvement de foule, inondations
  - Réseaux sociaux
- Attention
  - Si la visualisation peut aider à la compréhension d'un phénomène, elle peut introduire un biais et en altérer l'interprétation?
  - → Recherches fondamentales sur l'interaction H-M

L'apprentissage automatique : l'arbre qui cache la forêt

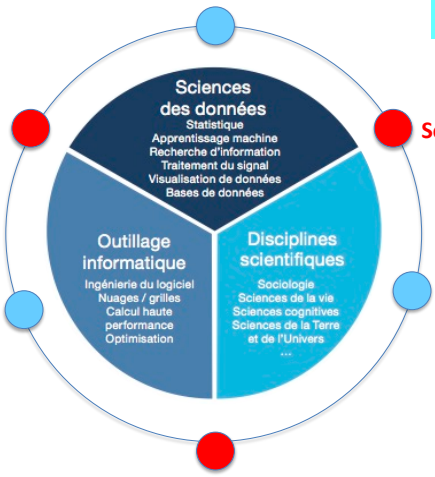


- Un outil d'extraction de connaissances et de reconnaissances de formes très performant
  - Nécessitant des millions d'exemples annotés qui vont servir pour l'entrainement
  - Basé sur des algorithmes très complexes et très gourmands en ressources de calcul
- Mais n'intervient qu'en bout de la chaîne de valorisation des big data
  - Acquisition des données
  - Qualification, annotation, nettoyage des données
  - Intégration des données
  - Stockage, indexation, tri, accès optimisé
  - …
- Pas pertinent pour tous les besoins
  - La corrélation de certaines données peut se faire par de "simples jointures" ou quelques opérateurs analytiques ("skyline")
  - Un bon langage de requêtes suffit pour beaucoup d'applications

L'émergence de nouveaux métiers en science des données

Trois grands domaines structurants

Trois métiers d'expertise



Expert en Science des données

Expert en Ingénierie des données

Expert des Sciences Appliquées

Expert en Génie Logiciel

Scientifique d'une discipline

Trois métiers d'intermédiation

Formateur Expert en Données

Le défi Mastodons

CNRS – DEFI MASTODONS

Objectifs

- La complexité croissante du monde ne peut plus se satisfaire d'une approche scientifique unique
  - éclatement de la connaissance
  - pluralité des savoirs sur des mêmes faits sociaux et sur leur articulation
- 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.
- Produire des concepts et des solutions qui n'auraient pu être obtenus sans coopération entre les différentes disciplines

Indicateurs de Suivi

- Pérennité de la coopération
- Publications communes
- Co-encadrement de thèses
- Plateformes de test et d'expérimentation
- Colloques/journées de dissémination
- Montage et soumission de nouveaux projets
- Synergie entre projets

Quelques chiffres

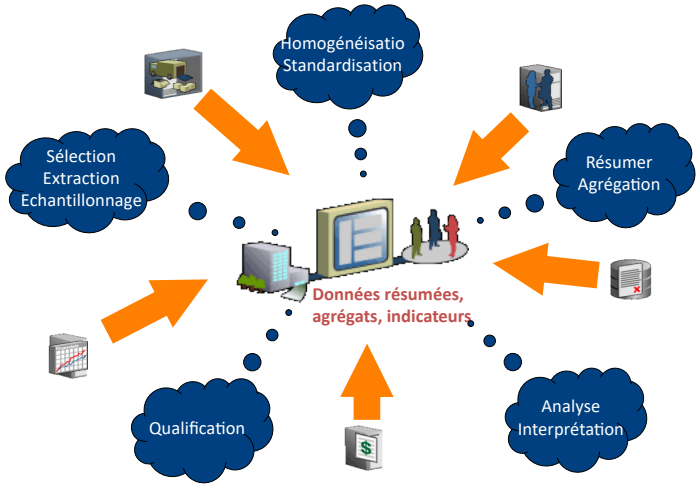
- **Défi lancé en 2012**
  - 50 projets retenus sur 120 soumis
- **Budget**
  - Près de 3,5 M€
  - Montant alloué/projet/an 20 à 120 K€
- **Degré d'implication des unités**
  - Plus de 100 labos impliqués, couvrant les 10 instituts du CNRS
  - Plusieurs EA universitaires associées
  - Plus de 400 chercheurs impliqués
- **Organismes impliqués hors CNRS**
  - INRIA, INRA, IRSTEA, INSERM, CEA, ONERA, Universités, Ecoles

Thématiques couvertes par les AAP

- **Collecte, stockage et indexation de données massives**
- **Hétérogénéité, interopérabilité, intégration, partage des données**
- **Calcul intensif sur des grands volumes de données, parallélisme dirigé par les données, optimisation**
- **Extraction de connaissances, datamining et apprentissage**
- **Préservation/archivage des données (animation)**

- **RI agrégative/résumé, sémantique et raisonnement**
- **Visualisation de grandes masses de données**
- **Qualité, protection et sécurité des données**
- **Problèmes de propriété, de droit d'usage, droit à l'oubli**
- **Consommation d'énergie, environnement, recyclage…**

Quelques problématiques communes aux différents projets



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, intégration, fusion de données
- Biologie santé
  - Génome, séquençage, phénotypage
  - Données médicales
- Réseaux sociaux
  - RI, analyse d'opinions, santé
  - Crowdsourcing

Bilan à 6 ans

- **Une communauté interdisciplinaire en formation sur les thèmes des Big Data et Science des Données**
  - **Organisée dans un GdR d'animation (MaDICS, plus de 600 adhérents en déc. 2015)**
- **Une production scientifique interdisciplinaire**
  - **Plusieurs publications signées par des chercheurs de différentes disciplines**
  - **Plusieurs co-encadrements de thèses**
- **Nouvelles initiatives de recherche issues du défi**
  - **Défi Imag'In sur le traitement d'images**
  - **Le PEPS FaSciDo sur les fondements des sciences de données**
  - **Le PEPS Astro-Informatique sur les sciences de données en astrophysique**
- **Levier pour initier d'autre projets ou réseaux de compétences**
  - **Plus de 10 projets ANR**
  - **2 actions COST (Europe)**

Le GdR MaDICS

- **Actions actuelles**
  1. **Apprentissage, optimisation Large-échelle et calculs distribués (ATLAS)**
  2. **Qualité des données scientifiques (ARQUADS)**
  3. **Environmental Acoustic Data Mining (EADM)**
  4. **Graph Data Mining (GRAMINEES)**
  5. **Imagerie Hyperspectrale (Imhyp)**
  6. **Masses de données en astronomie et astrophysique (MAESTRO)**
  7. **Préservation des données scientifiques (PREDON)**
  8. **Reproductibilité des expériences d'analyse de données scientifiques (ReProVirtuFlow)**
  9. **Raisonner sur les données (RoD)**

http://www.madics.fr

Quelques lectures