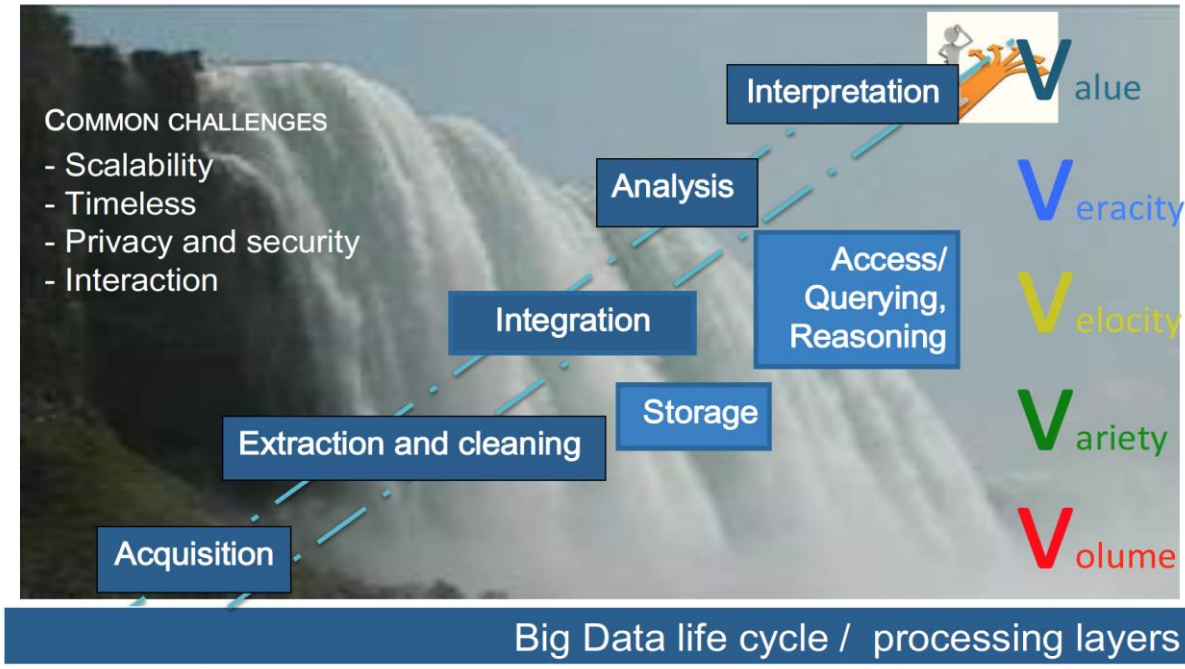
BIG DATA @AMU

C. Diaconu CPPM

Aix*Marseille université

Data is everywhere



inspired by "Big Data and Its Technical Challenges, Communications of the ACM, July 2014, vol 57, n°7", © H.V. Jagadish et all. 7

Big Data:

- Vitesse : Digitisation et disponibilité immédiate
- Variété : Des données hétérogènes
- Volume : Le volume dépasse les capacités de traite ment par les méthodes classiques.
- Veridicité: accès aux données authentifiées
- Valeur: données=blé (beaucoup)

Mais aussi:

- Visualisation
- Volatilité
- Vulnérabilité
- Vétusté
- Variabilité etc. etc.

Is the (beyond the) economy….

BIG DATA/ANALYTICS

Big Data to have big impact on UK economy: study

Analytics | Steve Evans | 04:46, April 5 2012

Tech's hottest topic can add billions to UK economy and spur the creation of new jobs, research claims

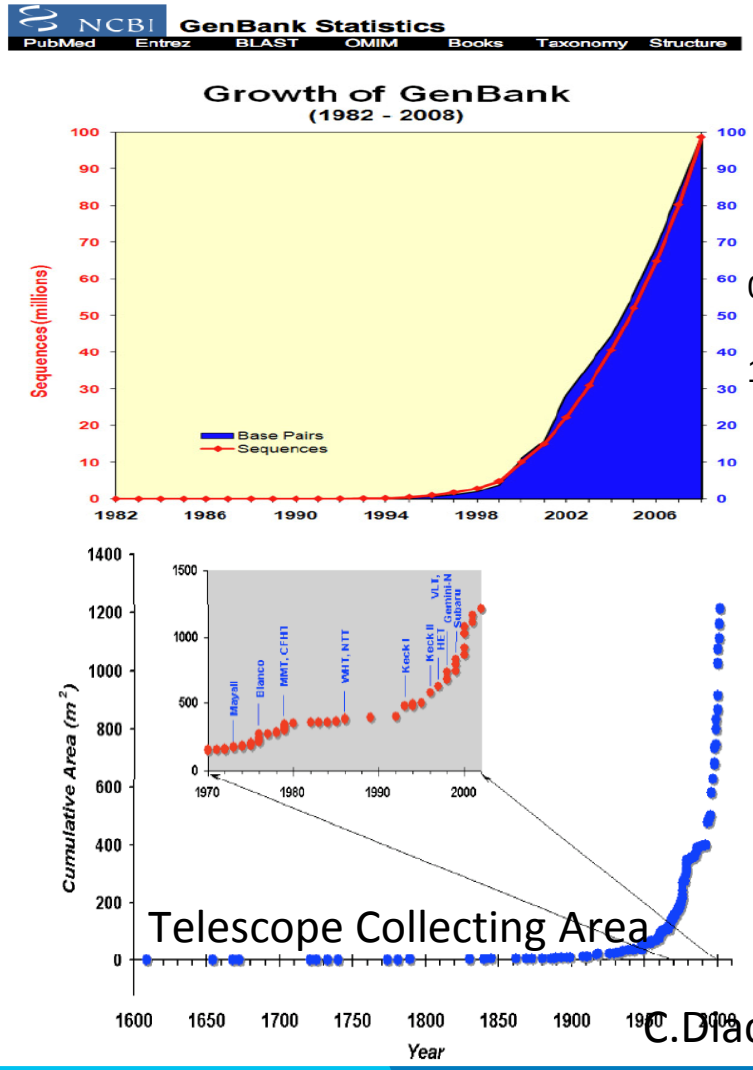
Big Data could potentially add £216bn to the UK economy while adding 58,000 new jobs by 2017, according to a new study.

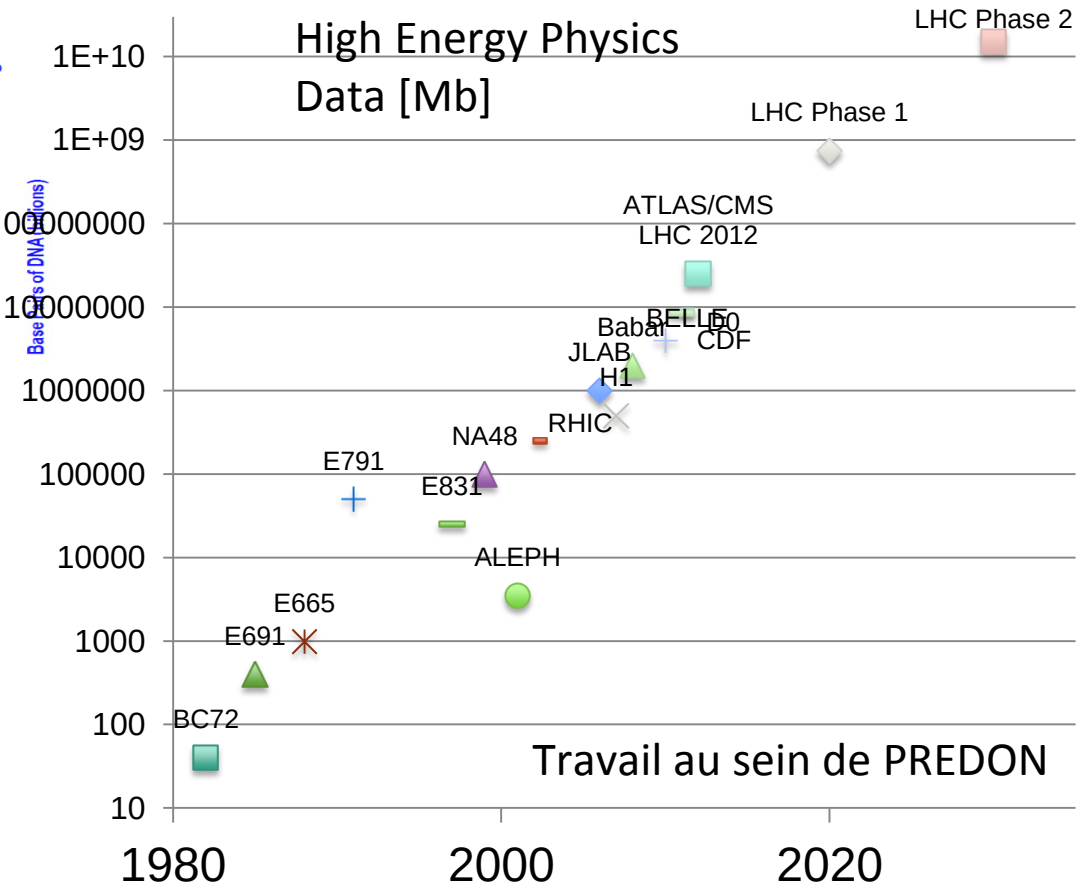
They studied 179 large companies and found that those adopting "data-driven decision making" **achieved productivity gains that were 5 percent to 6 percent higher than other factors could explain**.

« Big Scientific Data »

La recherche est « digitale »

- **Augmentation dramatique de la quantité/complexité des données**





© C.Diaconu

Comité de Pilotage « Big Data »

Coordination transverse des PR2Is
(Poles Interdisciplinaires et Intersectoriesl AMU)

- Identifier les acteurs AMU dans ce domaine
- Identifier les besoins
- Suivre les appels à projets, documenter et diffuser
- Stimuler les échanges et la construction des consortia
- Faire émerger des nouvelles collaborations
- Connexions au niveau national et international
- Connexion avec les PR2I, Amidex, AMU

Objectifs:
1) Cartographie des compétences (2014/2015)

2) Communication sur des sujets/axes affinées (2015-2016)

3) Construction de projets/consortia (2016-)

Sciences et Technologies
Cristinel Diaconu (CPPM) coord.
Christian Surace (LAM):
Thierry Artières (LIF)
Nicolas Ferré (Mesocentre)
Sebastien Fournier (LSIS)

Environnement
Joel Guiot (ECCOREV)
Jean-Pierre Bracco (LAMPEA

Energie
Mohamed Quafafou (LSIS)

Humanités
Patrice Bellot (LSIS)
Bernard Bel (LPL) (
Dominique Augey (Fac.Droit, AMU)
Fidelia Ibekwe-SanJuan IRSIC

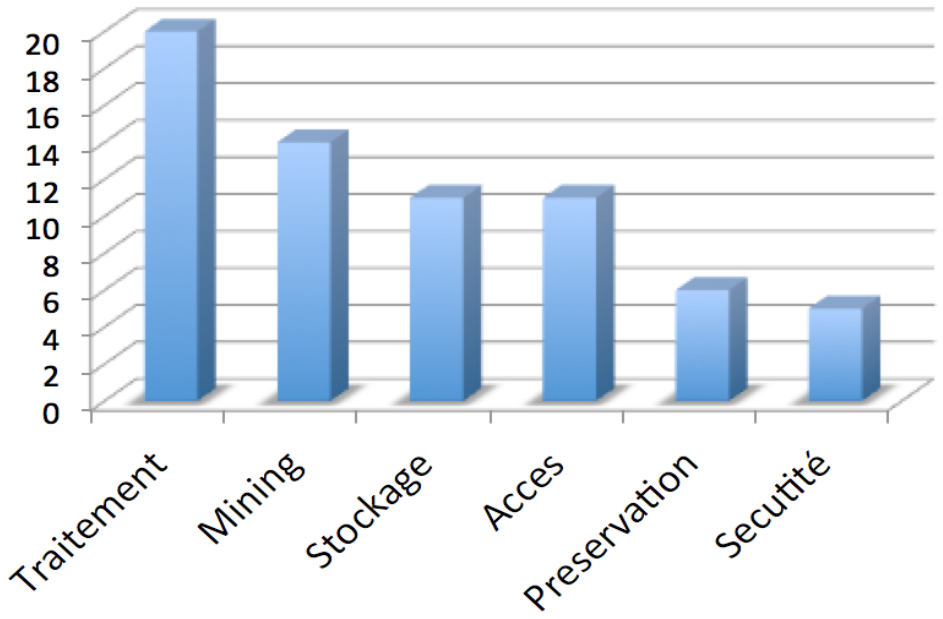
Santé
Frédéric Barras (LCB)
Jean-Charles Dufour (SESSTIM)
Christophe Béroud (Genetique)

AMU Recherche
Mossadek Talby (CPPM)

5/3/2016

AMU « Big Data »

Réunion « kick-off » November 24, 2014
~30 contact personnes dans les laboratories AMU



Synthèse des activités dans le domaine des grandes masses de données dans les laboratoires AMU

« GreyBook AMU » (evolving)

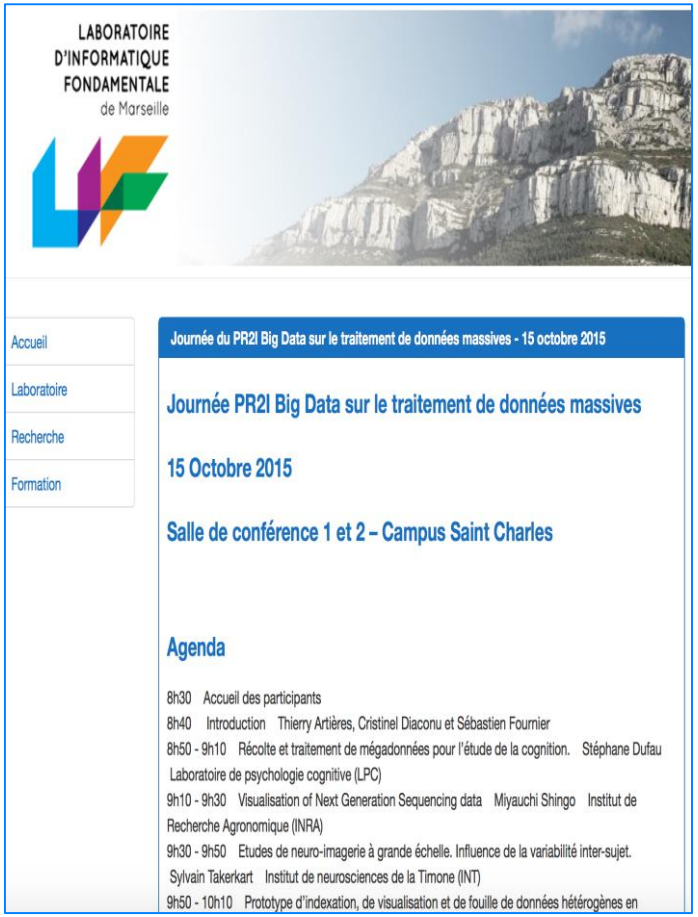
Opportunités: quatre axes

Data Mining	**Collaboration on methods and algorithms across disciplines** **Common projects on themes across disciplines** (examples: e-pub, biodiversity, health, social networks, urban analytics etc.) **Enhance the links with regional authorities**	**Common databases and their usage** (language, bioinformatics, etc.) **Develop common frameworks** **Scientific data preservation** **Heteregenous data integration**
Usage, ethics, legal issues	**Impact on management** **Ethics and society impact of big data** (example: personal genomics) **Legal issues of big data:** medial data, security, intellectual property	**AMU scientific data infrastructure**

Access Preservation

Infrastructure

Journées Thématiques « Big Data AMU »


LABORATOIRE D'INFORMATIQUE FONDAMENTALE de Marseille

Accueil
Laboratoire
Recherche
Formation

Journée du PR2I Big Data sur le traitement de données massives - 15 octobre 2015

Journée PR2I Big Data sur le traitement de données massives

15 Octobre 2015

Salle de conférence 1 et 2 – Campus Saint Charles

Agenda

8h30 Accueil des participants
8h40 Introduction Thierry Artières, Cristinel Diaconu et Sébastien Fournier
8h50 - 9h10 Récolte et traitement de mégadonnées pour l'étude de la cognition. Stéphane Dufau Laboratoire de psychologie cognitive (LPC)
9h10 - 9h30 Visualisation of Next Generation Sequencing data Miyauchi Shingo Institut de Recherche Agronomique (INRA)
9h30 - 9h50 Etudes de neuro-imagerie à grande échelle. Influence de la variabilité inter-sujet. Sylvain Takerkart Institut de neurosciences de la Timone (INT)
9h50 - 10h10 Prototype d'indexation, de visualisation et de fouille de données hétérogènes en


BIG DATA

ACCÈS
PRÉSERVATION
REPRODUCTIBILITÉ

VENDREDI 27 NOVEMBRE 2015
AMPHITEATRE DU CPPM
9H30

Comité d'organisation :
Cristinel Diaconu (CPPM, AMU)
Patrice Bellot (LSIS, AMU)
Christophe Béroud (Laboratoire de génétique, AMU)
Salima Benbernou (LIPADE, Paris Decartes)
Christian Surace (LAM, AMU)

CPPM PREDON Aix*Marseille université cnrs


BIG DATA

Enjeux, usages, éthique et droit du Big Data

Vendredi 25 mars 2016

à 9h30

Ecole de Journalisme et de Communication (EJCAM)

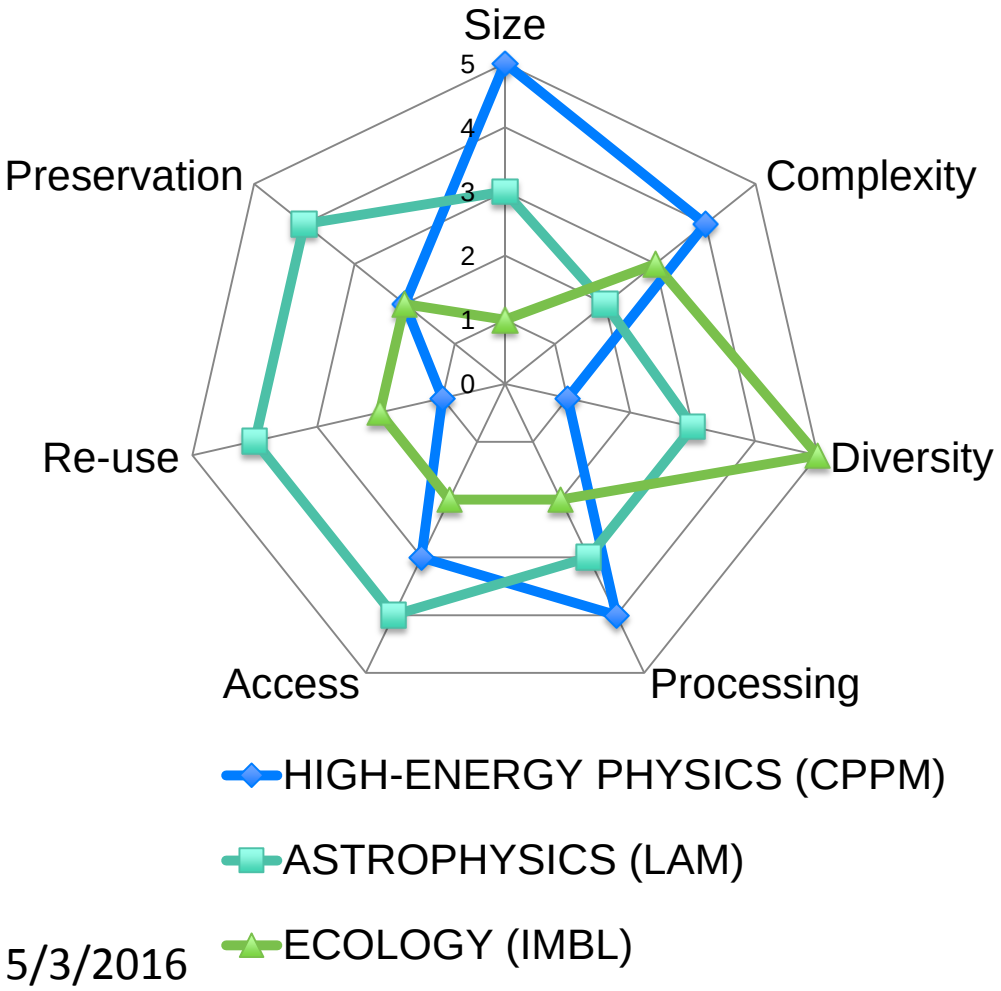
Amphitéâtre A

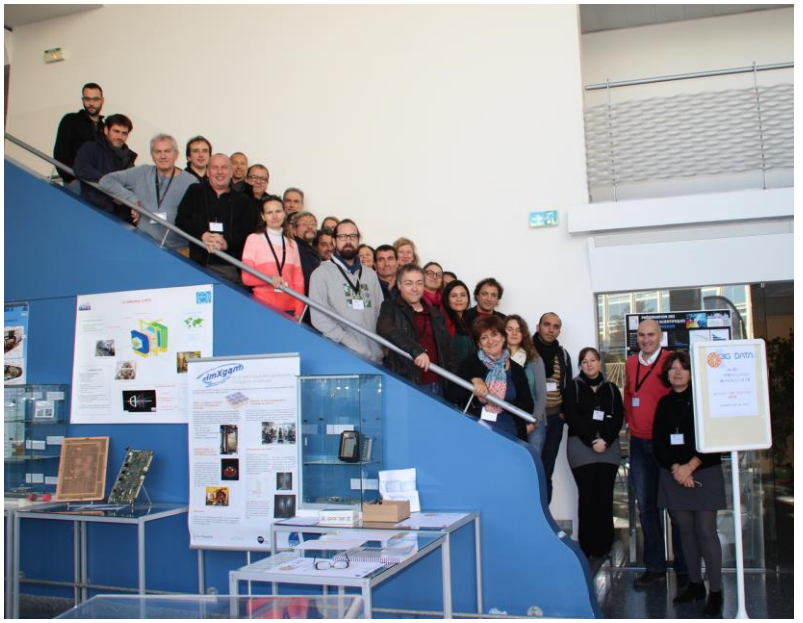
Comité d'organisation :
Françoise Bernard (IRSIC)
Cristinel Diaconu (CPPM)
Jean-Charles Dufour (SESSTIM)
Fidelia Ibekwe-SanJuan (IRSIC)
Alexandre Joux (IRSIC)

Aix*Marseille université cnrs

Journée sur les Infrastructures pour Big Data en préparation (juin)

Competences AMU





M³AMU

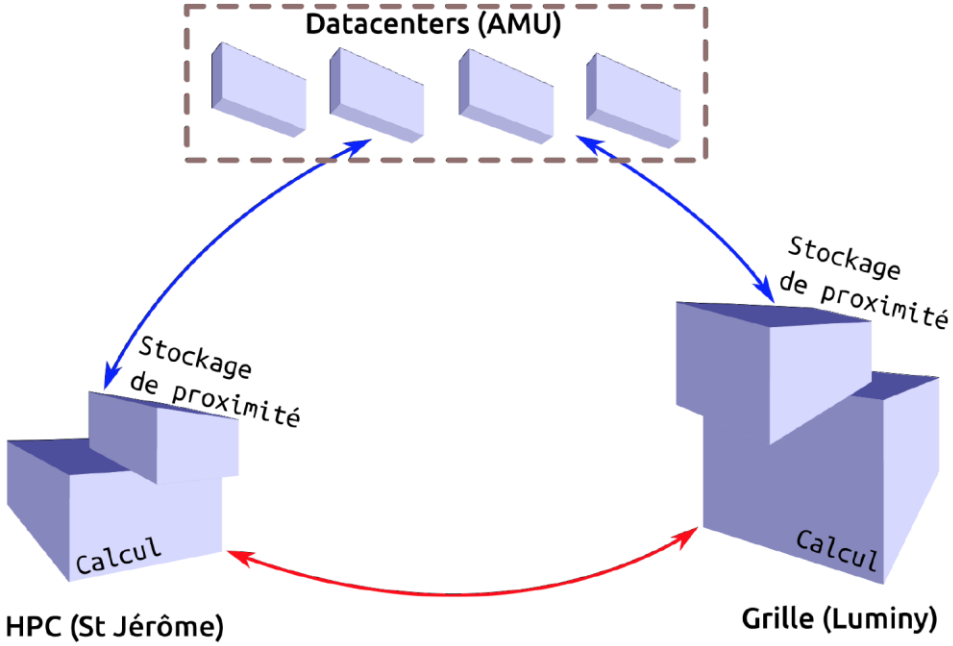
Mésocentre Multi-Modalités in AMU

HPC, grid, cloud, storage for scientific computing and Big Data

Project M3AMU funded by CPER/FEDER/CG13 (2016 – 1018) :

- Update + extension of the resources: +50 % more powerful
- Permanent storage close to the computing nodes
- Interface to the CPPM grid
- Cloud computing

Choosing an infrastructure for massive data: storage & processing?



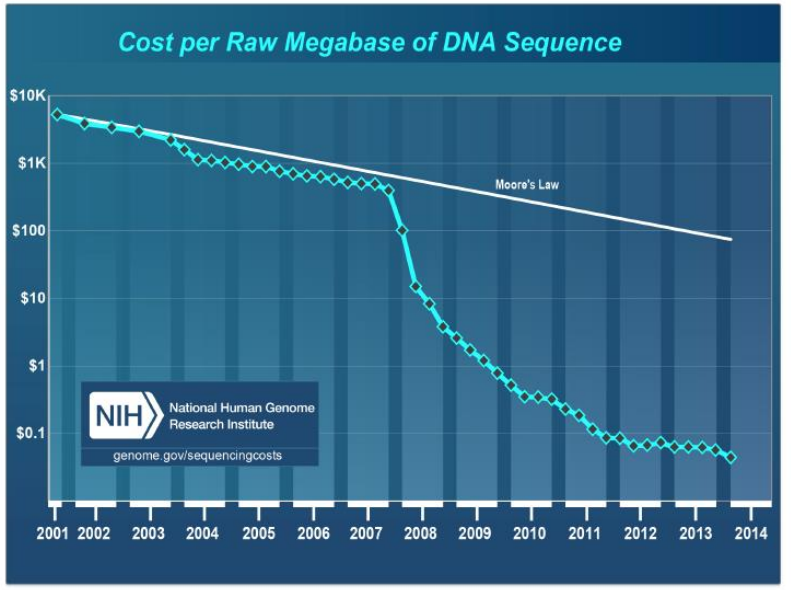
RHETICUS > 20 Tflops
1300 cores – 2.3 Tb RAM
+3.5 Tb of shared RAM

1958 cores on 98 servers
900 Tb stored on 18 servers
10 Gb/s network: LHCONE

Aix*Marseille université

BIG DATA

AMU Genetics: search for rare diseases



Le *"génome humain"* est connu

Contient **3,2 milliards** de paires de bases
- 1 page A4 contient ≈ 3 000 caractères
- 1 tome de 500 pages ≈ 1 500 000 bases
- 1 génome diploïde ≈ 2 150 tomes !

Il est possible de le séquencer pour chaque individu

Coût ≈ **3 000 € / individu**
- Pour avoir une bonne qualité : **30X**
- → **100 milliards de paires de bases / individu**
- → **64 500 tomes / individu**

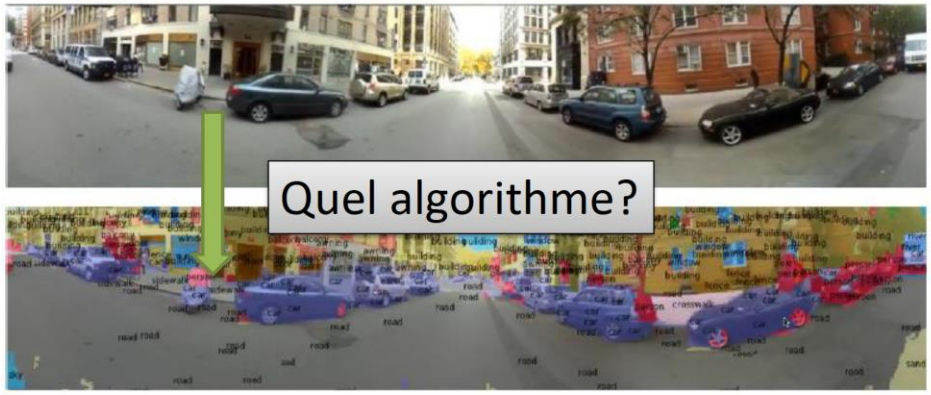
Find the disease-causing mutation

- **AMU is an EU leader in Human Genetics**
  - RD (research, diagnosis & treatment)
  - NGS
- **NGS data production**
  - Ion Proton (x2), NextSeq 500, PGM …
- **NGS data analysis (INSERM UMR_S910)**
  - High Performance Computing (UV2000 – 256cpu, 1To)
  - High Storage Capacity (1.2 Po)
  - Bioinformatics systems for data annotation and filtration
- **International networks**
- **Different teams**
- **Valorization**
  - BRCA Share™ (Quest Diagnosis / INSERM)

Professeur Christophe BEROUD
Laboratoire de Génétique Moléculaire
Hôpital TIMONE Enfants

SEQSTIM · CRS · TAGC Inserm U1090 technological advances for genomics and clinics · UMR_S910 Génétique Médicale et Génomique Fonctionnelle · Inserm · Aix*Marseille université

Apprentissage automatique (deep learning)

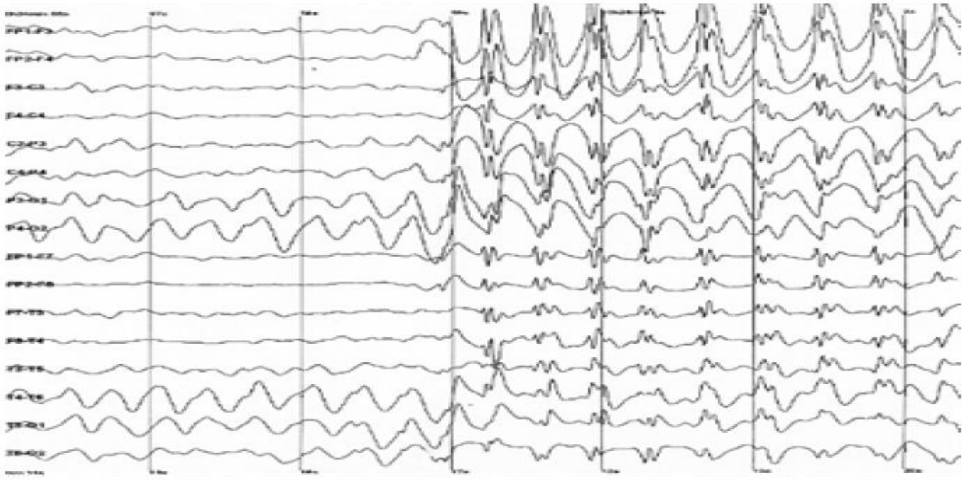


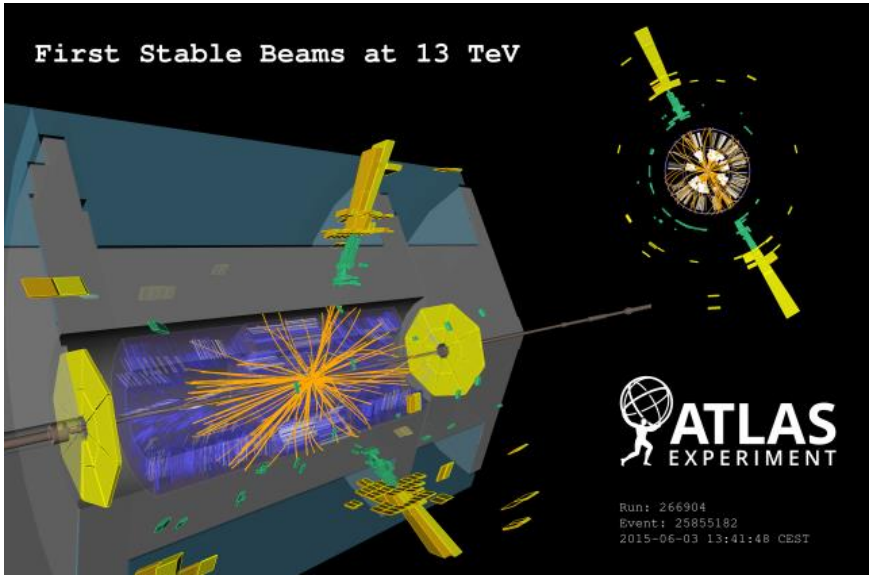
[Farabet et al., IEEE PAMI, 2012] Annotation automatique de scènes visuelles

- Champs d'applications de + en +nombreux
  - Taches perceptives : images, vidéos, parole, musique, gestes
  - Robotique et systèmes autonomes
  - Recherche d'information, traduction automatique, analyse de sentiment, analyse de réputation
  - Recommandation et personnalisation, web advertising…
  - Intelligence Artificielle

Physique des particules au LHC?

Apprentissage profond pour la recherche de phénomènes physiques nouveaux à partir des données massives collectées par l'expérience ATLAS au Grand Collisionneur de Hadrons (LHC) au CERN. (sujet de thèse - CD, Y. Coadou, T.Artieres-LIF)





Big Data en SHS

Des projets et des infrastructures

Patrice Bellot LSIS

Ethique, droit, usags, journalisme

Aix*Marseille université

Faculté de Droit et de Science Politique Aix*Marseille Université

Is 'Big Data' beautiful ?

COLLOQUE MÉDIAS014
12 DÉCEMBRE 2014 | AIX-EN-PROVENCE
Faculté de Droit et de Science Politique - 3, avenue Robert Schuman
9h00 - 18h00
Espace Cassin - Amphithéâtre Favoreu

Des pratiques à la limite du droit

RUBRIQUES EN CONTINU OPINIONS BLOGS IMAGES

LOGEMENT

Avec Airbnb, ma petite entreprise de location ne connaît pas la crise

8 minutes de lecture

Julie Conti
Publié dimanche 9 novembre 2014 à 12:05.

PARTAGE

De plus en plus d'appartements sont proposés sur le site internet Airbnb, de particulier à particulier. Il s'agit parfois de résidences principales, mais aussi de logements loués uniquement aux voyageurs

Des formations "data" toutes récentes

Le Data-Journalisme fait son entrée dans le programme de SciencesCom

La première formation au data-journalisme commence début janvier.

Pin it

SciencesCom est une école qui s'adapte toujours aux innovations dans les secteurs de la communication et des médias. C'est pour cela que le 16 janvier 2016, l'école va lancer la première formation en data-journalisme auprès de ses Masters 1.

Le data-journalisme (journalisme de données) est une évolution du secteur, et prend de plus en plus d'ampleur. Il s'agit de l'exploitation et de la mise en images de données sous des formats plus ou moins structurés. Les étudiants de SciencesCom vont donc avoir une formation afin de comprendre et connaître le data-journalisme, l'Open Data, les méthodes pour sa mise en oeuvre et surtout la révolution journalistique engendrée.

Aix*Marseille université

PREDON http://predon.org

Préservation des données scientifiques

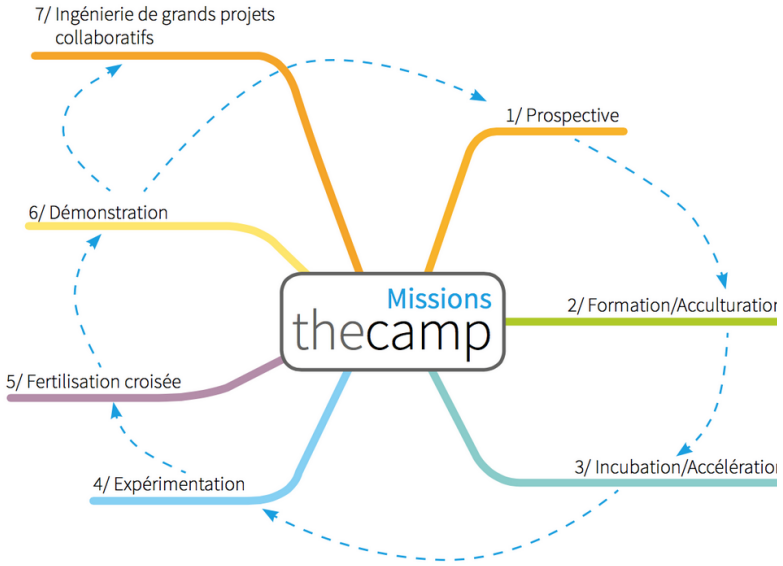
	Volume données	Complexité	Diversification des sources	Structuration au niveau international	Algorithmes et methodologie pour la preservation
IN2P3 HEP	+++	+++	+	++	+
INSU, IRD Astrophysics Earth Sciences	++	++	++	+++	++
CINES INS2I IT, Algorithms, workflows	+	++	+++	+	+++

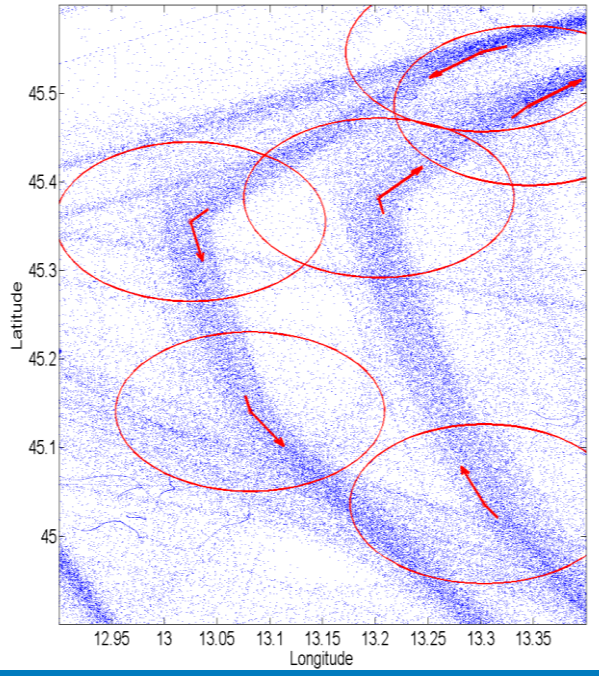
Projet « Mastodons/Big Data » de la MI/CNRS
Action dans Madics (GDR Big Data)

MaDICS
Masses de Données, Informations et Connaissances en Sciences
Big Data - Data Science

Big Data et les entreprises

Tres large potentiel de coopération
Poles compétitivité
-SCS Solutions Communicantes Sécurisées
-Pôle Mer Méditerranée (Sealab)
-Pôle Eau
-thecamp (Arbois, F. Chevalier)
.....





Semaine AMU entreprises

DU 2 AU 6 NOVEMBRE 2015
MARSEILLE & AIX-EN-PROVENCE

Aix*Marseille université

Table Ronde « Big Data »

- Les défis des données massives
- Les données massives dans la recherche
- Les données massives dans l'industrie
- La formation pour le "Big Data »

Jeudi 5 Novembre 2015

Faculté de Droit et de Science Politique
3 Avenue Robert Schuman, 13100 Aix-en-Provence
Salle des Actes

Participants:

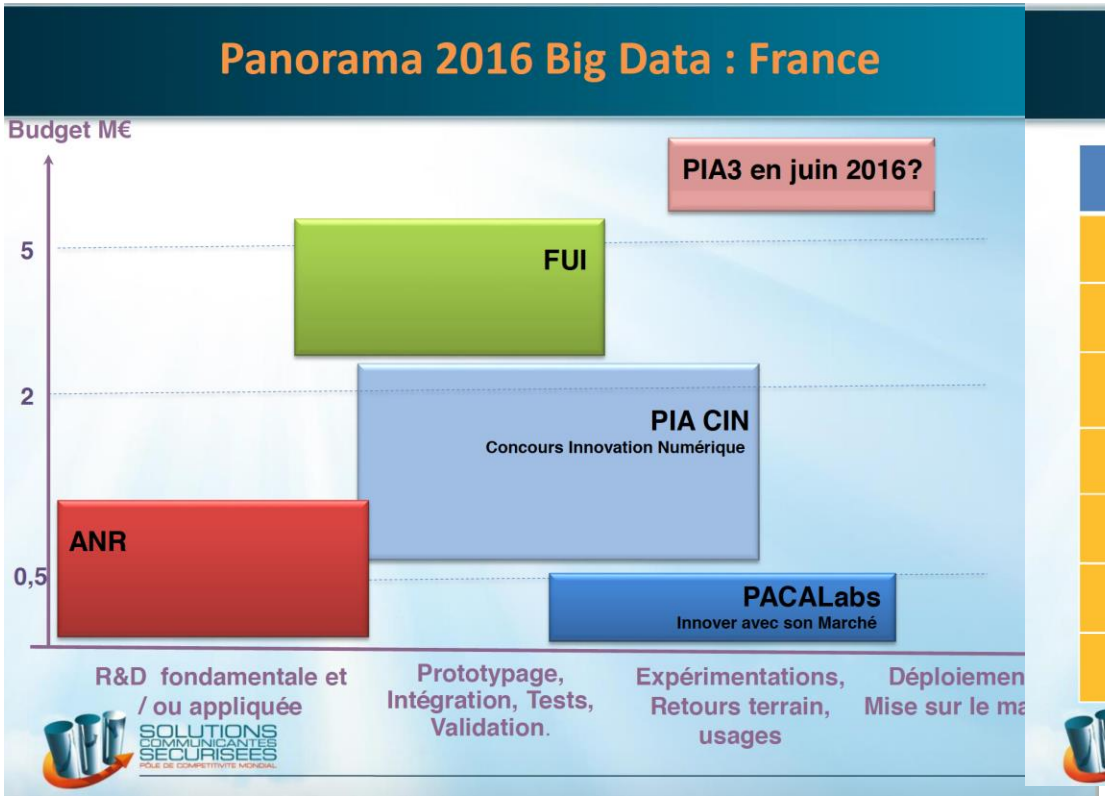
Thierry Artières, Laboratoire d'Informatique Fondamentale, AMU et Ecole centrale de Marseille, «Big Algorithmes»
Dominique Augey, GREQAM Laboratoire d'économie quantitative, Faculté de Droit et de Science Politique AMU
Christian Béroud, Laboratoire de Génétique Moléculaire, AMU , «Big Data en génétique»
Stéphane Claisse, Directeur Adjoint du Pôle Mer Méditerranée, « Big Data: The Sea Challenge »
Nicolas Ferre, Institut de Chimie Radicalaire, AMU, « Le mésocentre AMU: des ressources hardware pour le Big Data »
Fabien Finucci, Délégué Régional Orange, « Big data: une réalité opérationnelle »
Carine Nourry, Directrice de l'Ecole Aix-Marseille School of Economics, «Former nos étudiants à relever les défis du big data.»
Jean-Alexis Tanchou, PDG SEARIS, « Big Data: data mining, visualisation de la conaissance »

Animateur:

Cristinel Diaconu, Centre de Physique des Particules de Marseille, AMU

www.univ-amu.fr

Programmes de financement recherche+entreprises



Résumé
Appels Projets France 2016

	FUI	PIA CIN	PACALabs
Consortium Minimal	2 E +1 ACA	1 E	1 E
Budget Moyen	2 à 5 M€	0,5 à 3M€	0,1 à 0,5 M€
Aides Ent	SUB 25 à 45%	SUB+AR	AR ou PTZI 60%
Aides ACA	SUB 40% CC ou 100% CM	-	SUB 40% CC ou 100% CM
Thématiques	NON	OUI	OUI
Label Pôle	OUI	Option	NON
TRLs	4 - 7	5 - 8	7 - 9

SOLUTIONS COMMUNICANTES SECURISEES

Paysage national

Recherche dynamique

Prise de conscience du potentiel enseignement/formation

Consultation, structuration

Coordination interdisciplinaire

http://www.madics.fr/

http://www.faire-simple.gouv.fr/bigdatasante

Formation

Selon le ministère de l'innovation et de l'économie numérique, on estimait début 2014 à 300 000 le nombre de data scientists nécessaires à l'Europe dans les années à venir.

- Institut de l'entreprise «**Faire entrer la France dans la troisième révolution industrielle : le pari de l'innovation »**
- http://www.institut-entreprise.fr/sites/default/files/docs/Big-data.pdf

À ce jour, il existe encore peu de formations françaises en ce domaine :

- le Mastère Spécialisé « Big Data : Gestion et analyse des données massives (BGD) » de Telecom Paris-Tech, a ouvert à la rentrée 2013,
- Mastère Spécialisé en Big Data a été lancé par l'Ensimag (Grenoble INP) et l'EMSI Grenoble (GEM) pour la rentrée 2014.
- AMU School of Economics: Master Big Data (ingénieurs « Big Data »)

Par comparaison, plus d'une vingtaine d'universités américaines ont lancé ou devraient lancer des formations big data.

- L'Université Columbia (New York) 'IDSE (Institute for Data Sciences and Engineering) un nouveau diplôme intitulé « Certification of Professional achievement in Data Sciences ». probabilités et statistiques, algorithmes pour big data, machine learning et exploration des données.
- L'Université de Stanford délivre quant à elle un cours en ligne depuis 2013 orienté vers les big data : « Mining Massive Data Sets ».

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

The shortage of data scientists is becoming a serious constraint in some sectors.

Des formations "data" toutes récentes

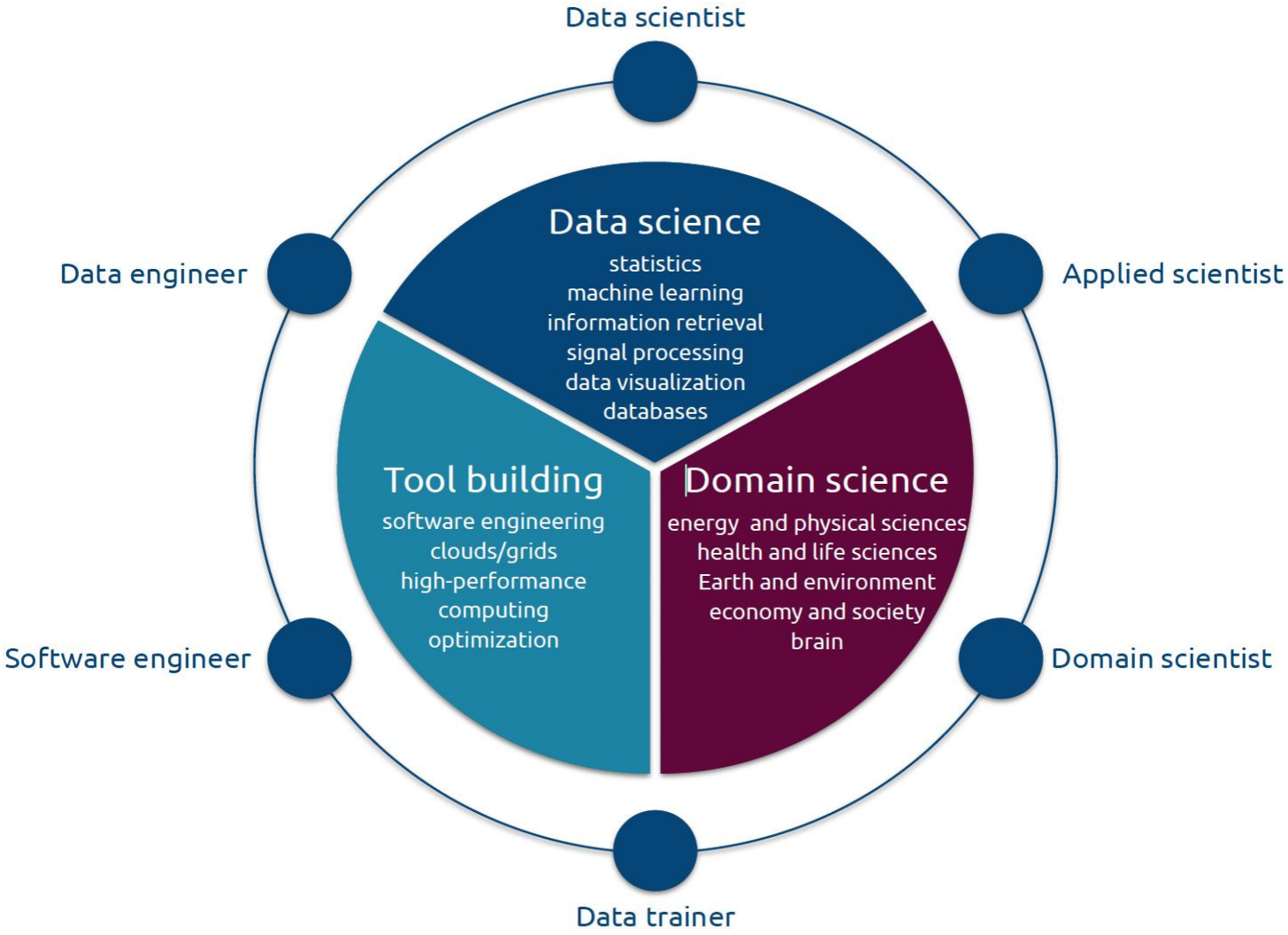
Le Data-Journalisme fait son entrée dans le programme de SciencesCom

La première formation au data-journalisme commence début janvier.

SciencesCom est une école qui s'adapte toujours aux innovations dans les secteurs de la communication et des médias. C'est pour cela que le 16 janvier 2016, l'école va lancer la première formation en data-journalisme auprès de ses Masters 1.

Le data-journalisme (journalisme de données) est une évolution du secteur, et prend de plus en plus d'ampleur. Il s'agit de l'exploitation et de la mise en images de données sous des formats plus ou moins structurés. Les étudiants

Training for the future



Enseignements AMU

Des coursus/cours dans la thématique (AMSE, Polytech etc.)

amse
école d'économie d'aix-marseille
aix-marseille school of economics

BIG DATA, NOW AND TOMORROW: SO MANY OPPORTUNITIES!

March 11th, 2016

Site Jules Ferry amphi B, 14 av. Jules Ferry, 13621 Aix-en-Provence

The recent evolution of ITs now allows to collect, store and analyze in real time large volumes of data of quite various types. The aim of this workshop is to show that these « Big Data » may be quite helpful for providing relevant answers to questions that arise in private companies, in public administrations as well as in academic research. The role of Big Data in health, banking, finance, media audience measurement, economic forecasting, marketing are among the topics that will be discussed.

Les technologies informatiques permettent aujourd'hui de collecter, de stocker et d'analyser en temps réel des données nombreuses et de natures très diverses. L'objet de cette journée est de montrer que ces « Big Data » constituent une ressource importante qui aide à répondre de façon pertinente à des questions se posant dans les entreprises, dans les administrations ou dans le monde de la recherche. Les thèmes abordés lors de cette journée seront donc divers : santé, banque, mesures d'audience des media, prévision conjoncturelle, marketing entre autres.

PROGRAMME BIG DATA DU MAGISTERE

Overview of the program :

Year 1 / L3 (starting in September 2015)

Big data : an introduction.
Big Data : an introduction to Hadoop and other tools for treatment and a
Collecting and structuring data (SQL, noSQL);
data cleansing with Python ; data visualisation (qlikview, tableau)

+ informatique (Python)

Year 2 / M1 (starting in September 2016)

Advanced SAS
Python for data scientists

Year 3 / M2 (starting in September 2017)

Data mining, text mining, association rules mining
Managing Big Data with SAS
Practical machine learning

Visite du Conseil d'Orientation Scientifique de l'AMU Septembre 2015, nos messages:

- Big Data is an opportunity for AMU
- Exploit and enhance the leadership
- Prepare new projects and collaborations

Recommendations

1) AMU big data infrastructure that the **project M^{3}AMU** aims to set up, should address the objective of offering a support to all the research teams that need to manage considerable amounts of data[…] this **project could be very relevant to AMU,** and it could produce high impact outcomes (in terms of research activities and research projects.

4) Another challenge related to educational issues is **to train and prepare people to the new profession of data scientist.** One of the main characteristics in this context is the interdisciplinary. AMU should work at jointly defining with various experts within its laboratories a **master on data science**. The significant and original traits of this master could be represented by the variety of data repositories and related applications developed at AMU, such as those related to genomics, astrophysics, physics, Social and Human Sciences, etc.

5) To strengthen the process of unification, **still more emphasis on interdisciplinary should be given**, in order to encourage projects that exploit a synergy between the several skills developed within AMU in relation to distinct disciplines

Next Step: AMU Data Institute? Réunion 3 mai 2015

Big Data Institut AMU

AP ANR: Instituts Convergence

plusieurs initiatives AMU, démarrage de la reflexion

AMU-AMIDEX Instiutute

Possibilité d'implementer des "Instituts" dans les prochaines AP

Deadline possibles Septembre? Ou 2017. Ou plus tard.

Eviter d'enclencher un "rush", préparer à l'avance (quoi? Comment?)

Superposition avec d'autres initiatives, harmonisation, ambitions, cohérence

Demande d'information "laboratoires" (24 mars 2016)

- Laboratoire / Directeur / Personnel total (permanent recherche, total, visiteurs/an-estimation).
- Thématiques de recherche aux laboratoires (grandes lignes et sous-thématiques ; par exemple : Physique, nanotechnologies, lasers etc.)
- Equipes de recherche liées aux thématiques big data (thèmes, mots clefs).
- Thèmes de recherche liés aux grandes masses de données.
- Programmes interdisciplinaires en cours : description rapide mais axée sur l'excellence scientifique (mentionner les thématiques, les coopérations internationales, rôles de coordination des chercheurs, "milestones", publications, prix etc.) Mentionner ici les programmes d'envergure nationale (Labex, Equipex etc.) et internationale.
- Liste des enseignements délivrés par le laboratoire dans des thématiques proches des sciences de données. Noms des enseignants, niveau, nombre d'heures, audience (nombre d'étudiants).
- Nom des personnes et équivalent temps plein qui pourrait justifier d'une participation à des projets de recherche dans un IC.
- Infrastructures liées à cette thématique dans le laboratoire.
- Liaison avec le monde socio-économique (contact entreprises, projets communs)

Projets qui pourraient être mis en œuvre au sein de l'IC :

- **Programmes de recherche interdisciplinaires possibles au sein d'une structure locale comme IC (thématique, laboratoire / équipes impliquées, thématique / opportunités, déroulement possible, besoins possibles en RH et infrastructure, commentaires sur la localisation, vision à long terme, impact sur la formation et l'enseignement ;**
- **Formation / Enseignement (un master interdisciplinaire sur la science des données est en discussion, pour réunir les initiatives plus ciblées et donner plus d'ampleur, tout en profitant de la large assise pluridisciplinaire de l'AMU) ;**
- **Infrastructure interdisciplinaire ;**
- **Projet liés au monde socio-économique ;**
- **Commentaires sur l'organisation, localisation**

Responses : une quizaine de laboratoires (niveau de détail variable)

Demande d'information : "Formations"

Demande d'information sur les formations « données massives »

Liste des enseignements délivrés dans des thématiques proches des sciences de données. Noms des enseignants, niveau, nombre d'heures, audience (nombre d'étudiants).

Projets qui pourraient être mis en œuvre au sein de l'IC :

- **Formation / Enseignement (un master interdisciplinaire sur la science des données est en discussion, pour réunir les initiatives plus ciblées et donner plus d'ampleur, tout en profitant de la large assise pluridisciplinaire de l'AMU) ;**
- **Enseignements en lien avec le monde socio-économique ;**
- **Commentaires sur l'organisation, localisation**

Draft Proposition Institut ~~Convergence~~ AMU

Argumentaire Big Data
- Contexte
- Paysage national
- Impact, potentiel

Recherche
- Capital recherche, excellence au niveau national, international
- Coopérations possibles

Enseignement
- Etat des lieux: lister les compétences/enseignements
- Convergeance et complémentarité

Actions/propositions
- recherche: projets interdisciplinaires
- enseignement: stages, formations, seminaires, animation etc.

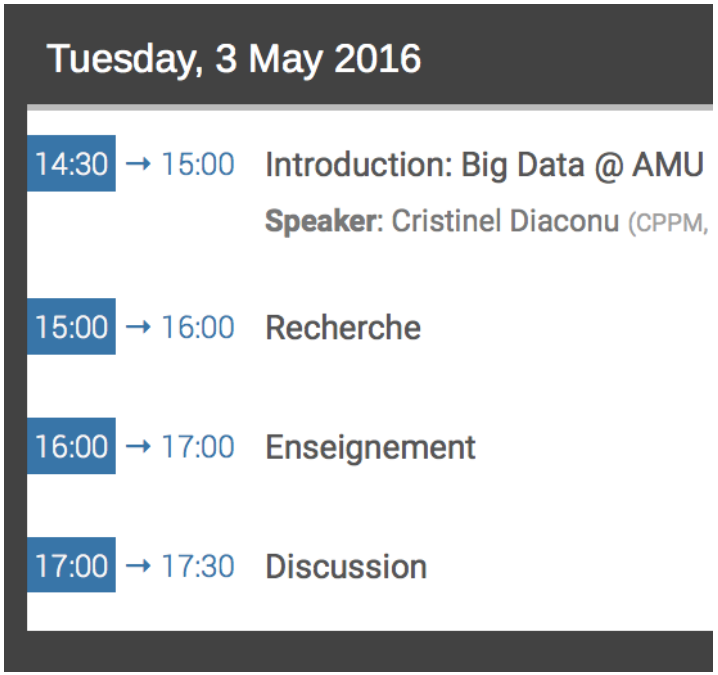
Structure et fonctionnement

Buts de la journée

Buts:
Prise de contact, collecte inputs
recherche, ***formation***
Préparation d'une repose à un AP "AMU Institute"
Drafter un document V000
Point de départ le draft IC circulé (agenda)

Organisation

Tour(s) de table
"Think Tank"
Minutes (volontaires?)



backup

Des *humanités numériques* - Big Data et SHS

Les données digitales sont fragiles

La capacité de stockage est physiquement dépassée depuis longtemps
Complexité, hétérogénéité, origine, reproductibilité etc.

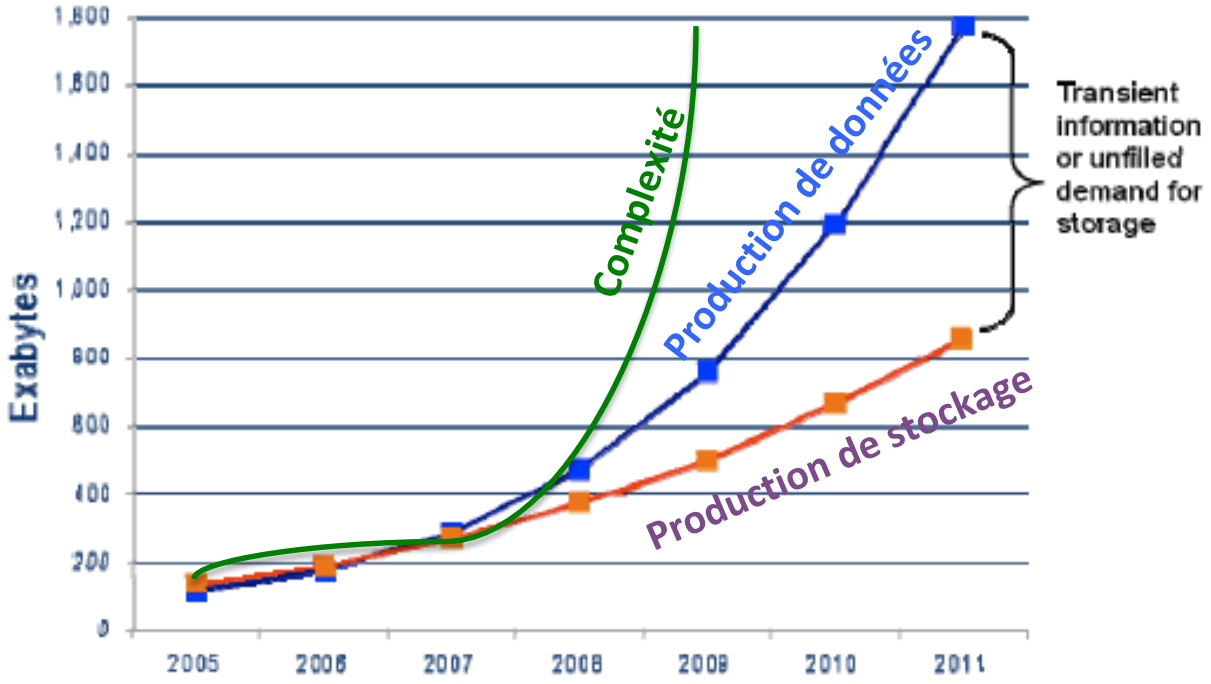


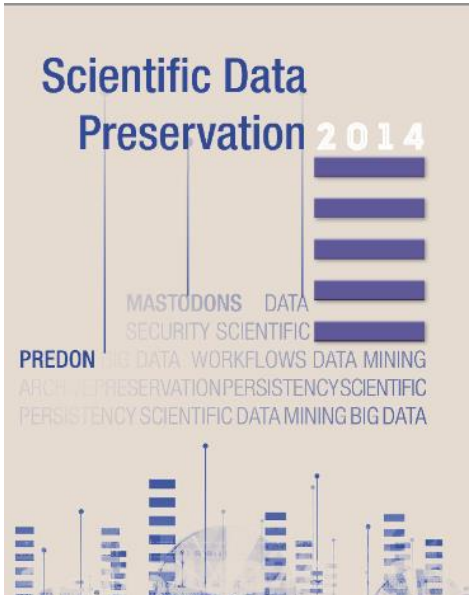
FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

Perte de données?

Big Data: Les 'V'?

- Valeur
- Veridicité
- Vitesse
- Varieté
- Volume
- **Vulnérabilité???**

Document PREDON 2014



Scientific case

9 DATA PRESERVATION IN HIGH ENERGY PHYSICS
C. Diaconu, S. Kraml

13 VIRTU
C. Su

17 CRYS TION: A WORLD-WIDE INITIATIVE
D. Ch

21 SATELLITE DATA MANAGEMENT AND PRESERVATION
T. Libourel , A. Laurent, Y. Lin

25 SEISMIC DATA PRESERVATION
M. Schaming

Scientific Case

Methodologies

31 WORKFLOWS AND SCIENTIFIC BIG DATA PRESERVATION
S. Benbernou, M. Lebbah

35 LONG
D. Bo

39 CLOU AND PRESERVATION
C. Cérin, M. Lebbah, H. Azzag

43 SCIENTIFIC DATA PRESERVATION, COPYRIGHT AND OPEN SCIENCE
P. Mouron

Technologies

49 STORA
J.- Y. M

51 REQUI IENTIFIC DATA AT CINES
S. Cou

57 VIRTUAL ENVIRONMENTS FOR DATA PRESERVATION
V. Be

Technologies

» http://informatique.in2p3.fr/li/?page=lettre&numero=27


n°27 Avril 2014

La lettre IN2P3 Informatique

Réseau des Informaticiens de l'IN2P3 et de l'IRFU

A LA UNE

Top départ pour la salle informatique « vallée » de Virtual Data !

Virtual Data est le groupe de travail du labex P2IO (réseau de tous les laboratoires du Sud de l'Ile de France impliqués dans la physique de l'infiniment petit, de l'infiniment grand et de l'étude des conditions d'apparition de la vie) dédié à la réflexion sur l'évolution des besoins informatiques des laboratoires du labex. Après des mois de réhabilitation d'un bâtiment de la vallée de l'université Paris-Sud, une longue phase de planification et de très nombreux allers et retours entre les salles informatiques historiques des laboratoires P2IO et la nouvelle salle « vallée » de VirtualData, un grand nombre de serveurs ont rejoint leurs nouveaux emplacements, concluant ainsi brillamment le déménagement d'une grande partie de l'infrastructure informatique de nos laboratoires de la vallée ! Ces machines utilisent maintenant presque totalement 18 racks sur les 30 prévus (pour une puissance totale de 400 kW) installés dans les 100 m2 de la première tranche de « Virtual Data », un projet porté par le Labex, mené à bien, dans les temps et dans le budget prévu, par un groupe de travail d'informaticiens des différents laboratoires de P2IO avec l'aide du service infrastructure du LAL.

Ce projet a pour but de regrouper les salles informatiques de huit laboratoires du campus au sein d'une plateforme commune, répartie sur une salle (...)

lire la suite

Interview

"Assurer notre excellence dans le domaine du calcul pour servir nos communautés scientifiques reste une priorité"

Ursula Bassler, directrice adjointe scientifique de l'IN2P3 en charge de la physique des particules et du calcul

lire l'interview

Stockage

Le service iRODS à l'IN2P3

Depuis plus d'une décennie, l'IN2P3 comme beaucoup d'autres acteurs du monde scientifique, est le témoin d'une explosion des quantités de données produites par les projets scientifiques qu'il soutient. La gestion de ces masses de données est un enjeu de première importance ; il faut à la fois assurer la

Développement

PYRAME

Pyrame, un framework de prototypage rapide pour systèmes online

Pyrame est un framework léger développé au Laboratoire Leprince-Ringuet (IN2P3/École Polytechnique) qui permet de prototyper rapidement et facilement des systèmes online de complexité arbitraire. Pyrame est basé sur une série de modules, gérant chacun un matériel spécifique ou une fonctionnalité système. Ils communiquent entre eux au

Collaboratif

eduroam

Accès nomade à Internet des personnels IN2P3 via « eduroam.in2p3.fr »

Le service eduroam vise à offrir un accès sans fil sécurisé à l'Internet pour les personnels des établissements d'enseignement supérieur et de recherche lors de leurs déplacements en France et à l'étranger. Les utilisateurs d'un établissement membre du

Préservation de données

Scientific Data Preservation 2014

PREDON : La

Imprimer

Agenda

EGI Community Forum, 19-23 mai - Helsinki

Le prochain Community Forum d'EGI (European Grid Infrastructure) se tiendra du 19 au 23 mai (...)

en savoir plus

HEPIX, 19-23 mai - LAPP

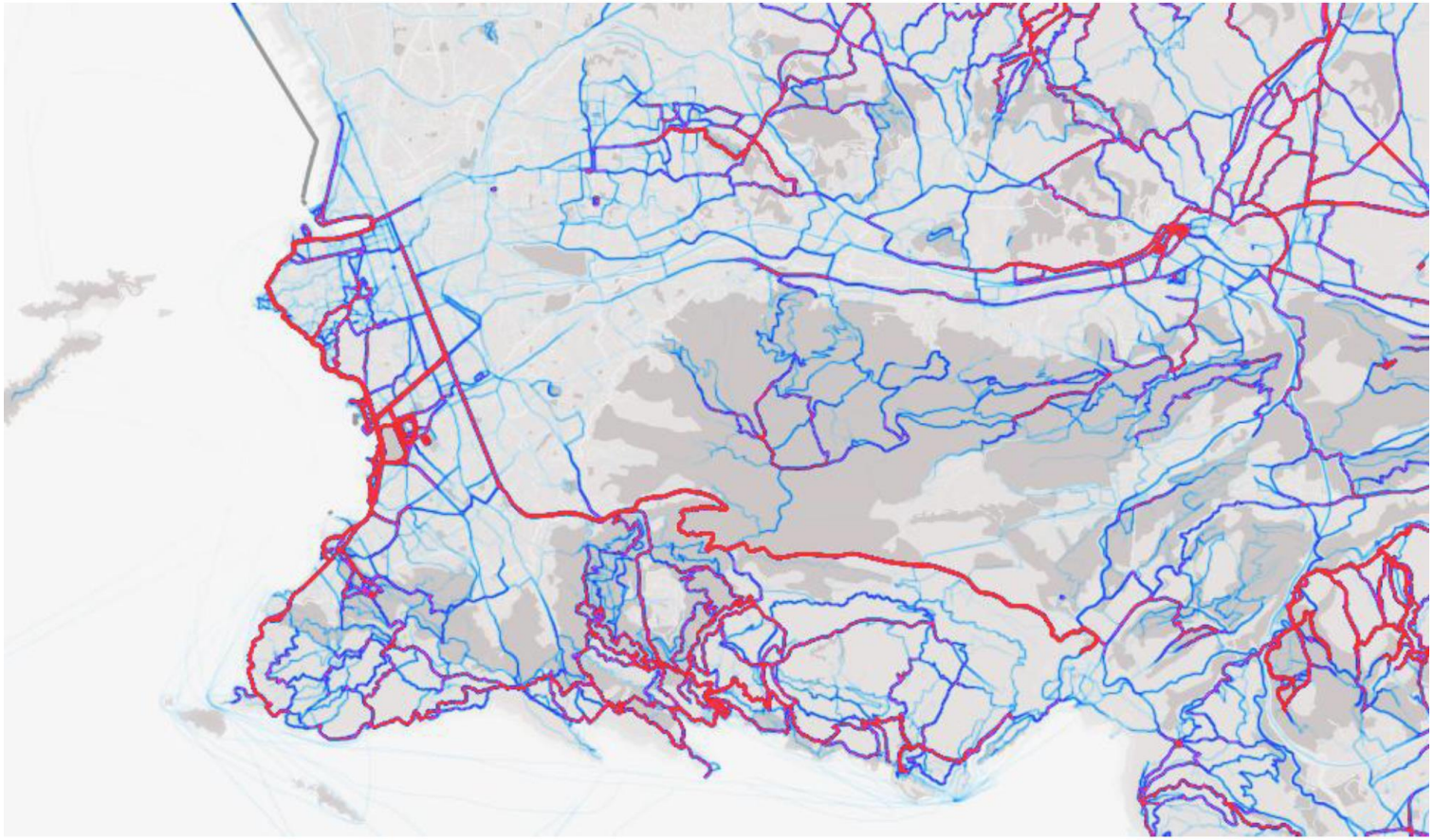
The Local Organizing Committee is pleased to announce that the registration and abstract (...)

en savoir plus

[normanDEV] Nouveau réseau

Etude des mobilités en géographie

Exemple : *Strava Heat Map* - Course à pied



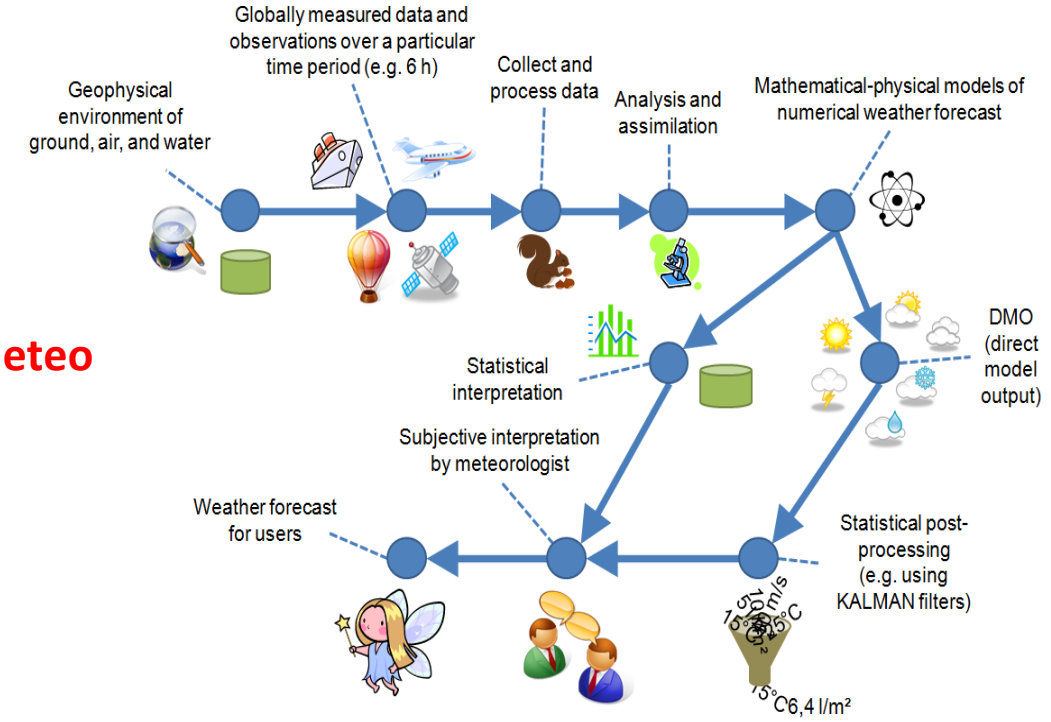
http://labs.strava.com/heatmap/#12/5.39034/43.25933/gray/run

« Big Data » Journalisme

Une aide aux choix éditoriaux

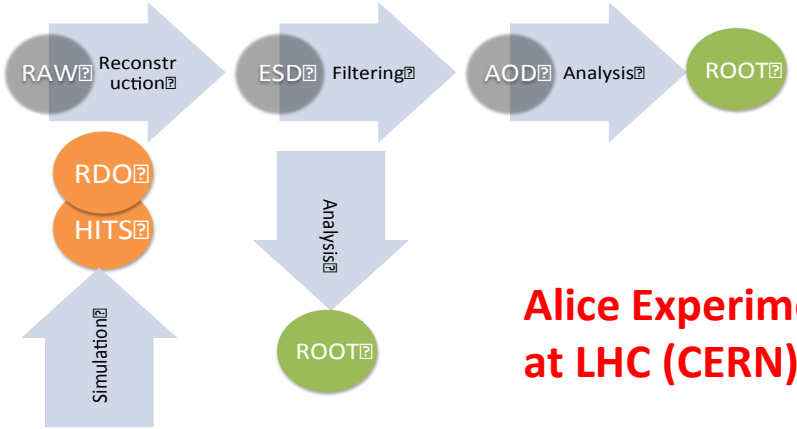
Mais aussi un domaine très nouveau pour les journalistes

Formats, workflows et préservation



Meteo

=



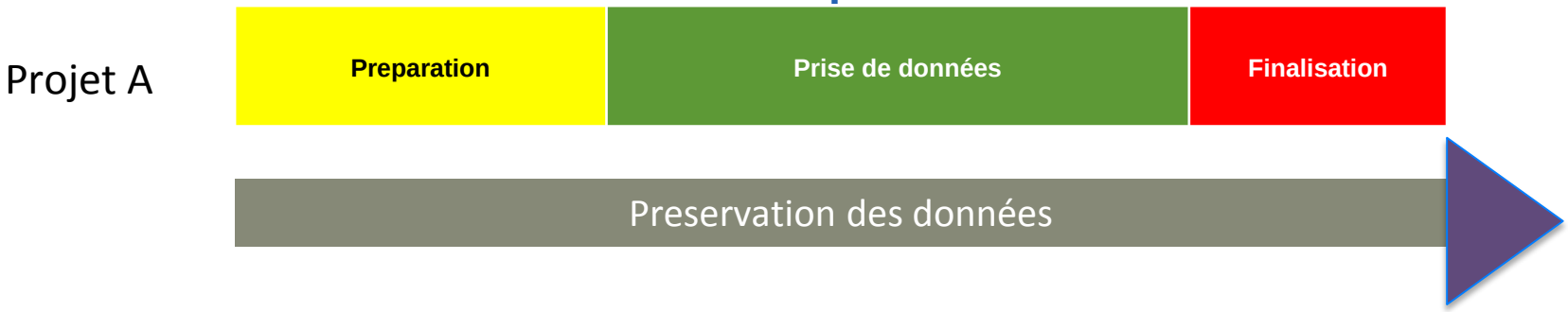
Alice Experiment at LHC (CERN)

Formats de données: standards?

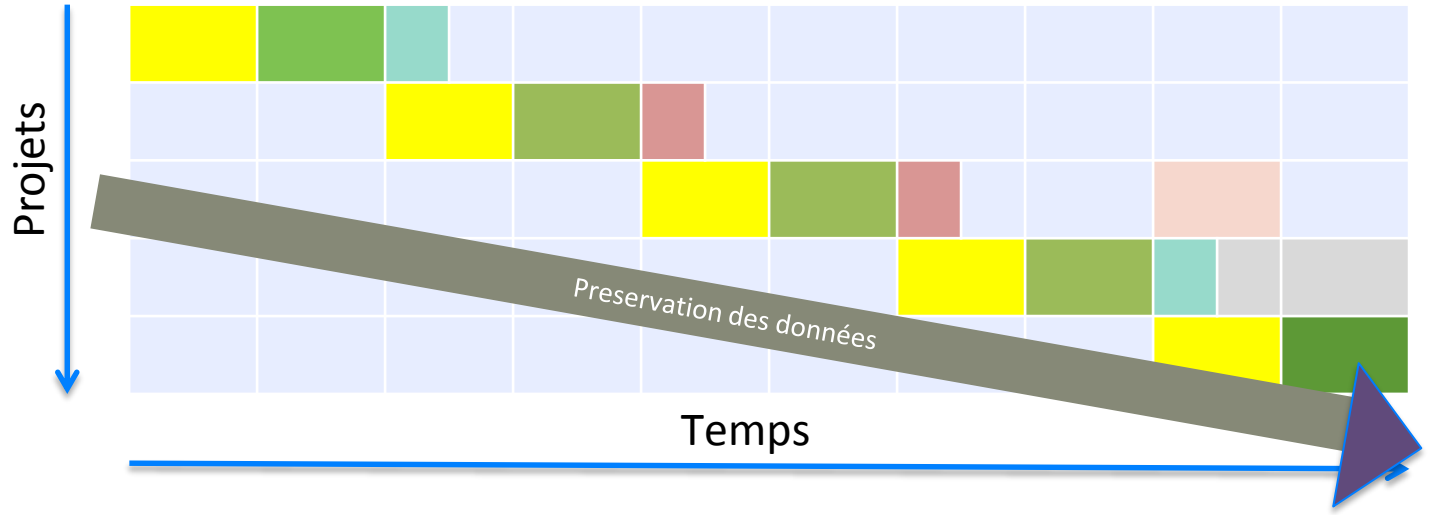
Similarité entre les disciplines

Approche théorique rigoureuse
Besoin et opportunité

Quand faut-il commencer à préserver?

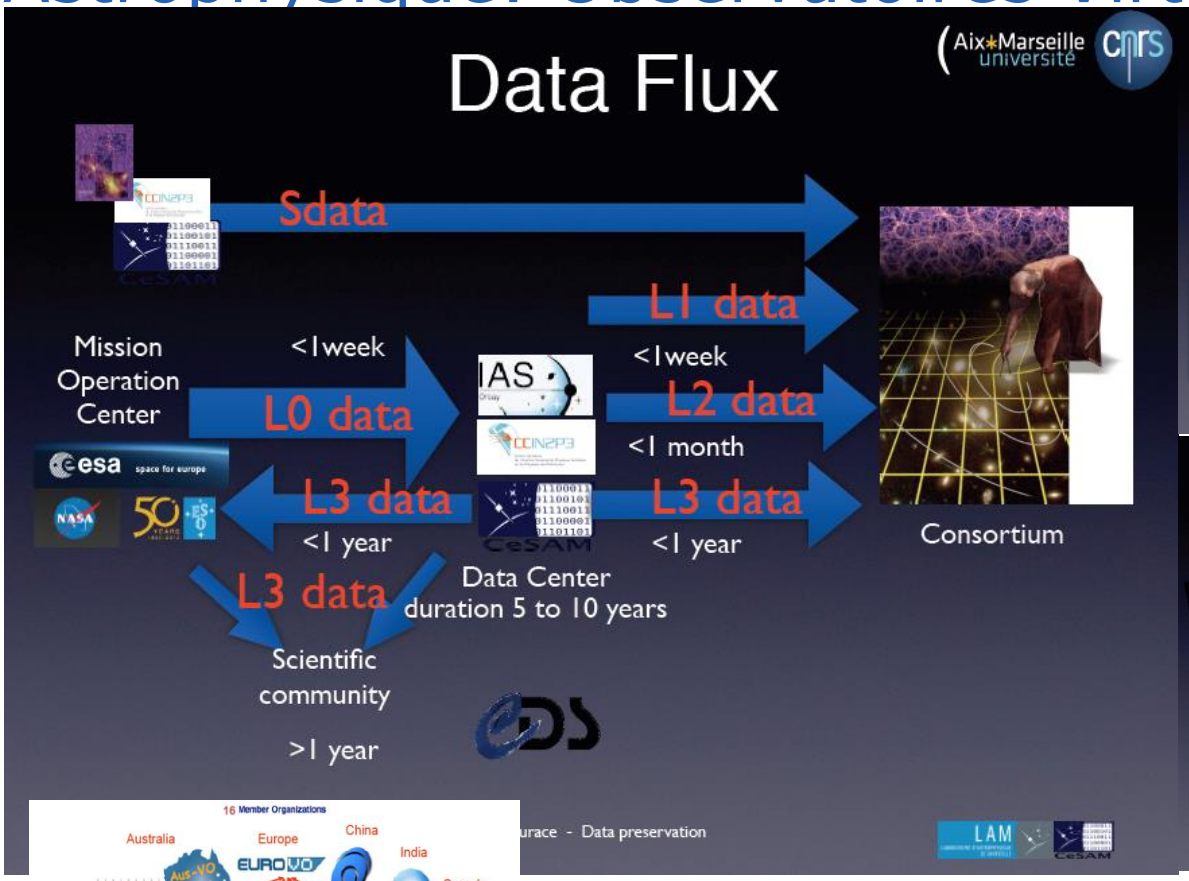


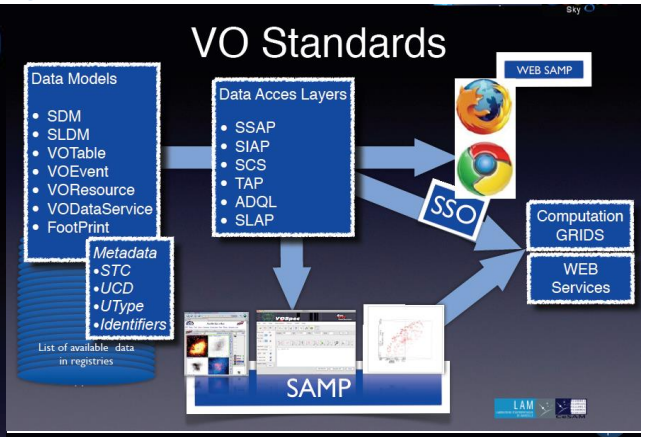
« Data Management Plan » doit inclure la préservation et l'accès à long tèrme



Programme cohérent de la préservation de données

Astrophysique: Observatoires Virtuels



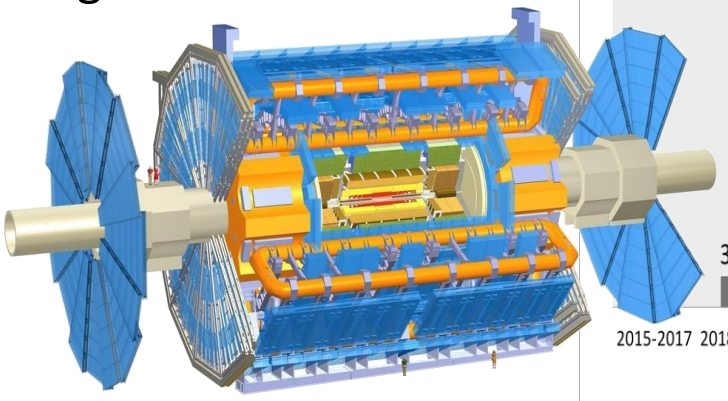


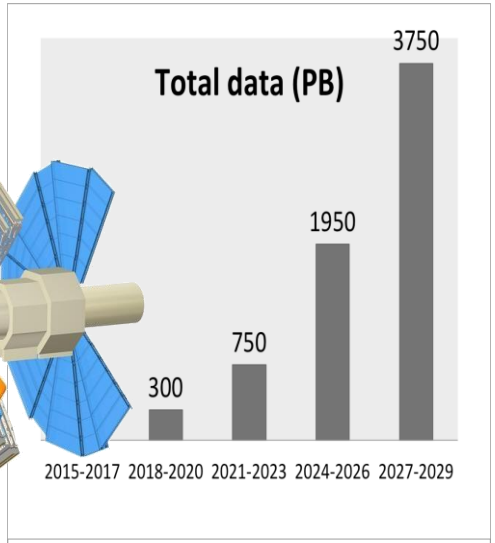
http://www.ivoa.org

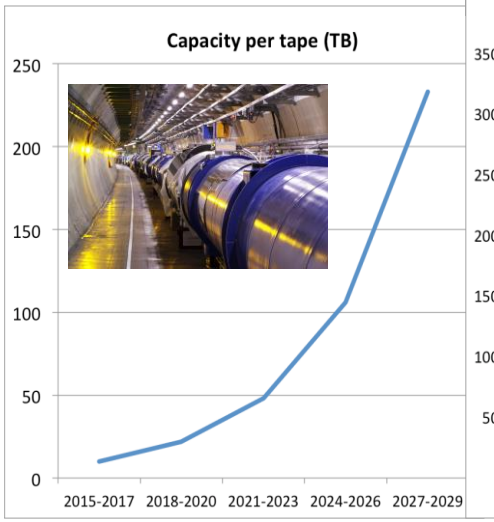
C.Diaconu

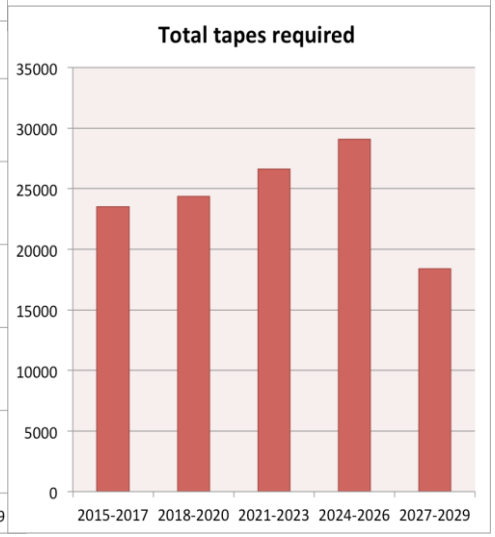
The problem is not (only) the storage

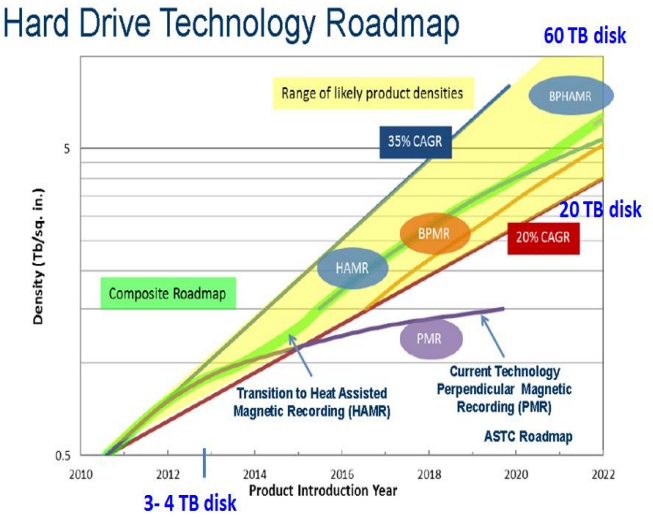
Example:
Large Hadron Collider











Méthodologies de traitement et fouille (« data mining »)

Compétences et points forts

- **Compétences dans les bases théoriques dans les méthodes de fouille, algorithmes, détection de corrélations, visualisation, classification de grande taille représentation, parallélisassions.**
- **Position de leadership au niveau national et international dans plusieurs domaines liés aux grandes masses de données (astrophysique, physique des particules, génomique)**
- **Développement d'outils et programmes (ex. bioinformatique), analyse de grandes masses de données.**
- **Panel de champs disciplinaires très complémentaire et favorisant les coopérations au niveau d'AMU avec un excellent potentiel au niveau international.**

Besoins

- **Collaborations souhaitées (recherche) entre fournisseurs de Big Data (biologistes, physiciens, chercheurs en sciences sociales, etc.), curateurs de données, statisticiens**
- **Formation et enseignement : master/licence, et aussi des formation pointues pour des chercheurs avancées.**
- **Plateformes d'échange données et logiciels, accès à des ressources communes.**

Opportunités et projets

- **Collaborations sur les méthodes et algorithmes : dimension spatiale des données (mobilités), représentations des données (sémiologie graphique), textes/ontologies.**
- **Montage des projets thématiques sur le Big Data incluant des domaines des sciences sociales, l'histoire, l'art, les biotechnologies, l'astronomie etc. Exemples de thématiques possibles/engagées : influence de l'e-publicité, traitement du langage, indexation des données sur la biodiversité méditerranéenne, biodiversité des habitats coralligènes, risque en finances et marchés, bio-diversité acoustique, données santé et gestion hospitalière, réseaux sociaux, géosciences, urbain analytics, renforcer nos relations avec des acteurs gestionnaires des territoires.**

Mise à disposition et préservation

Compétences et point forts

- **Projets d'envergure internationale dans la manipulation et l'accès aux grandes masses de données (Physique des particules, astrophysique, biologie).**
- **Participation à des projets de bases de données en biologie, pathologie, ressources pour la linguistique, microscopie, documents/Web, réseaux sociaux, données patrimoniales, géosciences ; bases de données rélationelles.**
- **Traitement systématique des données pluri-disciplinaires, compétences en stickage massif et pérenne. Participation à des projets dédiés à la préservation des données scientifiques au niveau national et international.**

Besoins

- **Plateformes de mise en commune des données des différentes disciplines qui pourrait stimuler l'échange et l'émergence de projets pluri- et inter-disciplinaires.**
- **Pour cela, une infrastructure matérielle importante (grappes de calcul, serveurs spécialisés, stockage) permettant de mettre en œuvre ces activités est nécessaire.**

Opportunités et projets

- **Traitement du langage et de ses bases cérébrales (LPL)**
- **Banques de données et des bases de données généralistes ou spécialisées, ainsi que des plates-formes logicielles donnant accès à des outils de bioinformatique et d'analyses d'images, des plus génériques aux plus pointus. Préservation pérenne de données scientifiques.**
- **Intégration de données hétérogènes, la mise en place de chaînes de traitement automatisées, la définition d'ontologies, la fouille de texte.**

Usages, droit et éthique

Compétences et points forts

- **Enjeux sociétaux du Big Data et de l'Open Data et aux dimensions épistémologiques et éthiques de leurs usages: collecte, du traitement et d'usages de grandes masses de données dans différents domaines tels que la santé, les médias et le journalisme, la publicité et le e-marketing, la recherche scientifique, etc**
- **Communication engageante numérique, économie numérique, intelligence économique et veille informationnelle**
- **Usages psychologiques et sociaux des medias sociaux en tant que « conscience collective virtuelle »**
- **Rajouter les compétences « sécurité (IML+LATP)**
- **Psicho-social (Mohamed)**
- **Sécurité (ici?)**

Besoins

- **La mise en commun des compétences dans les SHS dans ce domaine**
- **Formation et enseignement en éthique des données et préservation/archivage à long terme.**

Opportunités et projets

- **Etudes des transformations organisationnelles liées à l'émergence du digital.**
- **Enjeux éthiques et sociétaux de la génomique personnelle**
- **Propriété intellectuelle sur le code, droit et médecine (protection des données médicales), sécurité informatique, philosophie**

5/3/2016

Infrastructure et technologies

Competences

- **Infrastructures de calcul HPC (mésocentre) et Grille (Tier2 LHC, CPPM)**
- **Expertise dans l'utilisation massive de grand centres de calcul grille et HPC au niveau national et international.**
- **Accès à des technologies de pointe ; outils d'accès aux ressources, virtualisation, techniques de calcul sur le « cloud ».**

Besoins

- **Inter-connectivité, accès aux ressources distribuées, amélioration des bases d'infrastructure informatique dans certains laboratoires.**
- **Formation pour l'utilisation massive des grandes ressources informatiques.**
- **Besoins génériques en experts « data management », systèmes/infrastructures (réseau, Cloud, langages), expertises technologiques concernant les outils permettant de traiter le Big Data et d'intégrer des données issues de différents canaux et sources**

Opportunités et Projets

- **Projet de rapprochement entre le méso-centre et la grille du CPPM est en cours de développement. Lorsqu'elles sont disponibles, les ressources du méso-centre pourront être intégrées à Dirac, le logiciel de gestion de la grille du CPPM et participer ponctuellement au traitement très efficace de données massives.**

Préambule

- **Pôle de Recherche:** peu d'interactions directes avec l'enseignement.
- **Interdisciplinaire et intersectoriel:** Une recherche interdisciplinaire de qualité ne peut s'appuyer que sur des compétences fortes dans chacune des disciplines.
- **Le premier objectif** des pôles est l'animation, la mise en relation ciblée.
- **Le pôle est une aide** aux unités et ne doit pas être une contrainte pour les chercheurs.

5 Pôles de Recherche interdisciplinaires et Intersectoriels

- **Environnement : Nicolas Roche**
  Hommes, Milieux, Sociétés
- **Santé & Sciences de la Vie : Jean-Paul Borg**
  Innovations Biologiques & Biomédicales, Enjeux sanitaires et sociaux
- **Energies : Lounes Tadrist**
  Sources, Usages, Territoires, Politique et Sécurité Energétiques
- **Echanges et Dynamiques Transculturelles : Philippe Blache**
  Diversité des langues, des cultures, des économies et des sociétés
- **Sciences et Technologies : Philippe Delaporte**
  Ingénierie, Technologies Avancées et Sociétés

Motivations

- Structurer certaines activités de recherches autour de quelques thématiques ciblées et en s'appuyant sur des expertises fortes et des équipements sophistiqués (plateformes, réseaux …)
- Donner de la visibilité à AMU sur les thématiques choisies
- Être en capacité à répondre aux appels d'offres nationaux et internationaux
- Valoriser les compétences d'AMU sur ces thématiques auprès du tissu industriel (Pôles de compétitivité)

Les Activités Stratégiques en PACA

- Renforcer la dynamique d'innovation par **les Prides et Pôles de Compétitivité**
- Accompagner toutes **les entreprises** dans leur démarche d'innovation
- S'affirmer sur 2 grandes thématiques : L'économie Créative et l'économie de **la Méditerranée** durable
- S'inscrire dans une perspective d'innovation sociétale et territoriale déclinées en **cinq Domaines d'Activités Stratégiques (DAS) :**

1. Santé
2. Industrie du contenu numérique
3. Risques sécurité sûreté
4. Habitat durable
5. Mobilité intelligente

1. Transition énergétique-efficacité énergétique
2. Risques-sécurité-sûreté
3. Santé-alimentation
4. Mobilité intelligente et durable
5. Tourisme-industries culturelles-contenu numérique

Grands défis sociétaux de l'ANR

9 grands défis sociétaux

- Gestion sobre des ressources et adaptation au changement climatique
- Energie, propre, sûre et efficace
- Renouveau industriel
- Santé et bien-être
- Sécurité alimentaire et défi démographique
- Mobilité et systèmes urbains durables
- Société de l'information et de la communication
- Sociétés innovantes, intégrantes et adaptatives
- Liberté et sécurité de l'Europe, de ses citoyens et de ses résidents

Strengths

Data Mining	**Theoretical basis** **Large/complex data sets** **Tools and frameworks Complementarity**	**Data preservation** **Data-bases in many disciplines** **Massive and long-term storage**
Usage, ethics, legal issues	**Society impact of open/big data** **Epistemology and ethics** **Media studies**	**Large infrastructures** **Massive computing** **Cutting edge technology (virtualisation, cloud…) etc.)**

Access Preservation

Infrastructure

Needs

Data Mining	**Trans/Inter disciplinary cooperation** **Advanced and Student training programs** **Common centrally managed resources**	**Multi-disciplinary approach** **Common Platforms for expertise exchange and cooperation**
Usage, ethics, legal issues	**SHS expertise mutualisation** **Interdisciplinary cooperation** **Training** **New topics: security psicho-social and use impact**	**Inter-connectivité, distributed ressources , basic infrastructure** **Training** **Experts in data managment** **« Big data »-like expertise**

Access Preservation

Infrastructure

Données scientifiques

De plus en plus complexes
- Information riche, collectée par des capteurs versatiles

Encore plus vulnérables:
- modèle économique de la préservation à long terme quasi inexistant

Motivation scientifique évidente
- Recherche à bas coût, retour sur l'investissement


nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue

Archive > Volume 503 > Issue 7477 > News > Article

NATURE | NEWS

عربي

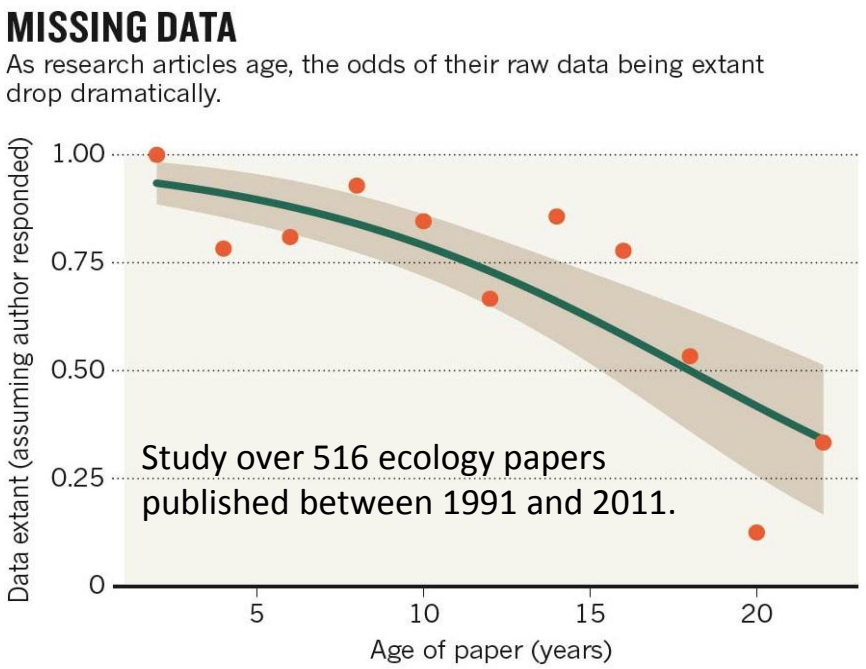
LHC plans for open data future

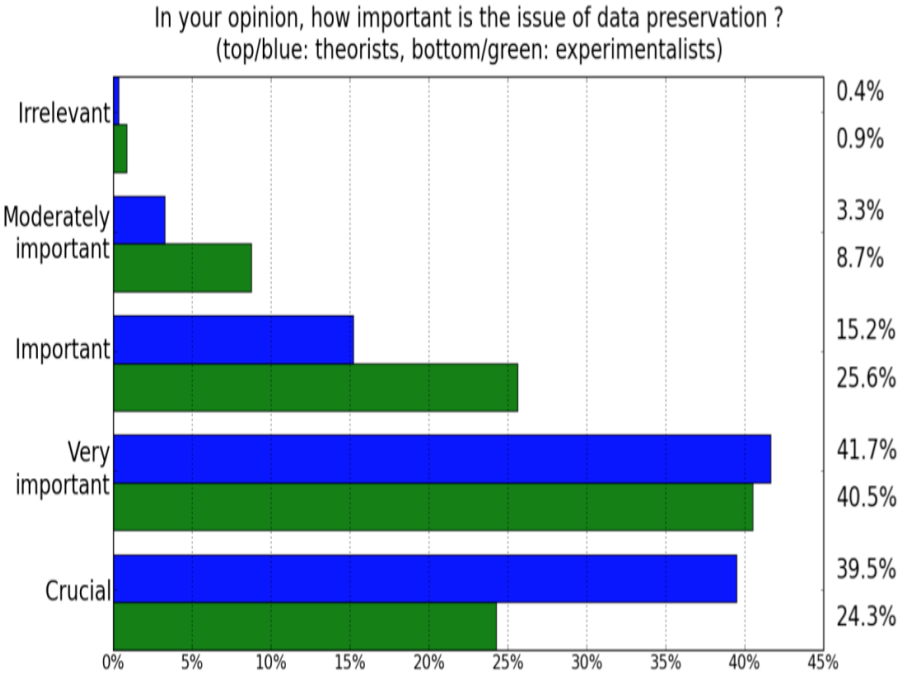
Researchers share results to keep them accessible.

Elizabeth Gibney

26 November 2013

"When the LHC programme comes to an end, it will probably be the last data at this frontier for many years. We can't afford to lose it."





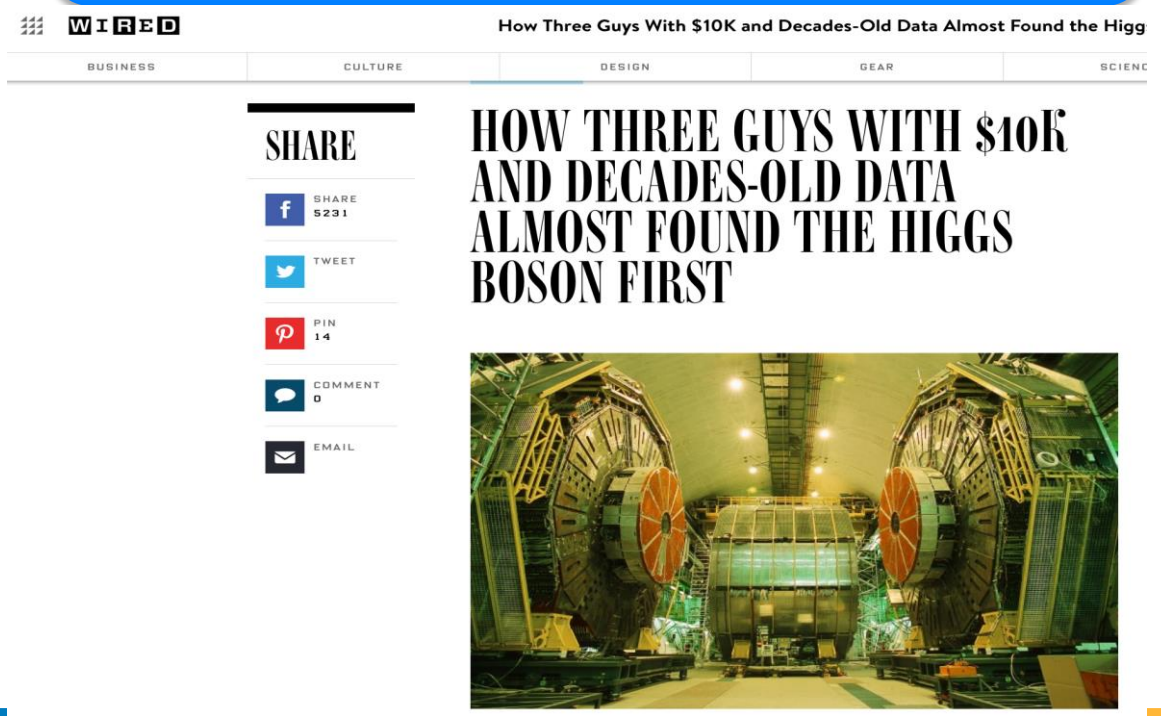
Les données scientifiques

Publications
Documentation
Donées (brutes+processées)
Meta-données
Workflows
Software
Diffuse knowledge
….more…

Challenges:
Complexité, couts

Approches:
Technologie,
Méthodologie
Organisation

Quel modèle de préservation pour les données scientifiques?

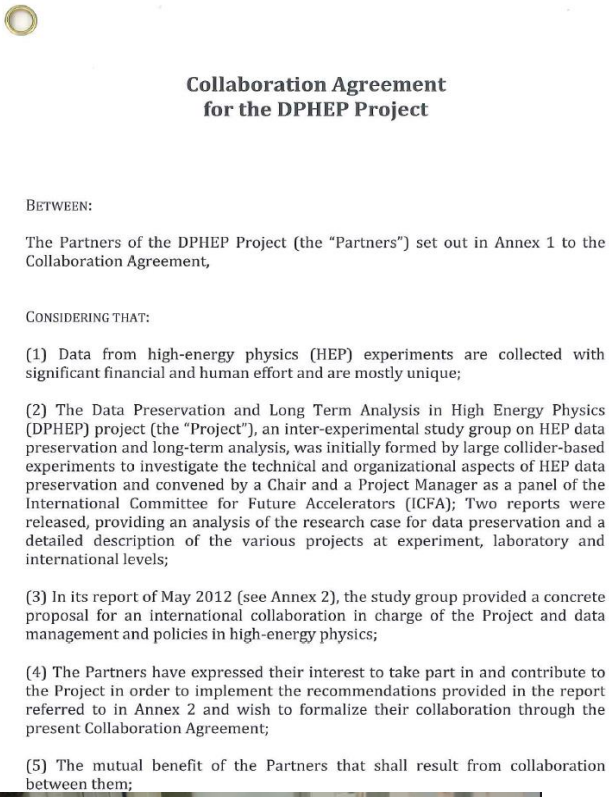


Structuration dans la physique des particules

DPHEP « Memorandum of understanding » signé par des agences de financement:

- **Suisse(CERN), France (IN2P3), Japon(KEK), Finlande (IPHY), Allemagne (DESY, MPI), Chine(IHEP), Canada(IoP)**

CERN: portal « open data » pour les données du LHC



Collaboration Agreement for the DPHEP Project

BETWEEN:

The Partners of the DPHEP Project (the "Partners") set out in Annex 1 to the Collaboration Agreement,

CONSIDERING THAT:

(1) Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique;

(2) The Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) project (the "Project"), an inter-experimental study group on HEP data preservation and long-term analysis, was initially formed by large collider-based experiments to investigate the technical and organizational aspects of HEP data preservation and convened by a Chair and a Project Manager as a panel of the International Committee for Future Accelerators (ICFA); Two reports were released, providing an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels;

(3) In its report of May 2012 (see Annex 2), the study group provided a concrete proposal for an international collaboration in charge of the Project and data management and policies in high-energy physics;

(4) The Partners have expressed their interest to take part in and contribute to the Project in order to implement the recommendations provided in the report referred to in Annex 2 and wish to formalize their collaboration through the present Collaboration Agreement;

(5) The mutual benefit of the Partners that shall result from collaboration between them;

First DPHEP Collaboration Board 2015