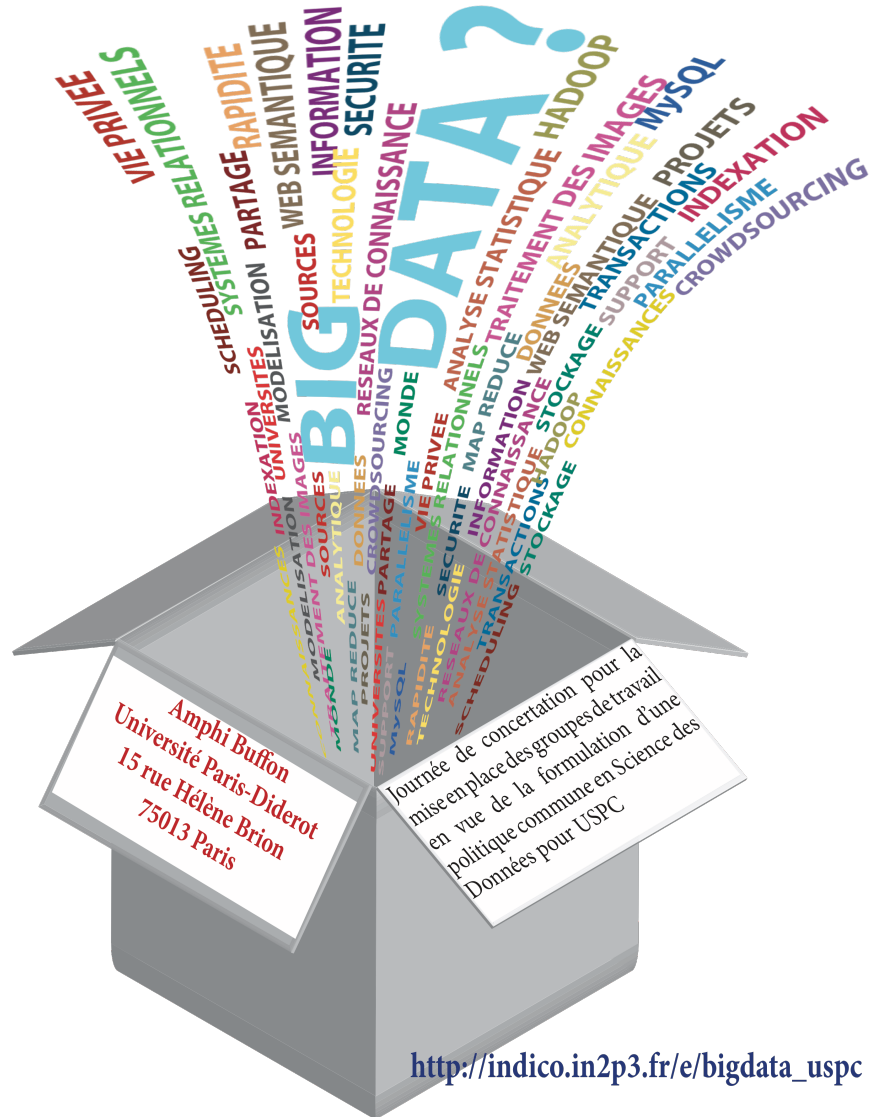


BIG DATA @ USPC

Lundi 30 novembre 2015

Remarques d'introduction

S.Katsanevas



http://indico.in2p3.fr/e/bigdata_uspc

C'est quoi Big Data ?

- Vitesse de calcul → Loi Moore x2 chaque 18 mois
- Capacité de stockage → Loi Kryder x2 chaque 13 mois
- Capacité du réseau → Loi Nielsen x2 chaque 21 mois
- Capteurs distribués → Smart Cities, Internet des Objets
- Analytique, origines académiques (liste non exhaustive...)
 - Math, Statistique, machine learning → découverte des lois dans des données
 - T. Hey : Théorie → experimentation → simulation → fouille des données - correlations, induction. Nouveau paradigme épistémologique ?
 - Analyse d'image
 - Reconstruction de données complexes
 - Objets connectés.
 - ...
- Impact: environnement, énergie, emploi, urbanisation, santé, vieillesse, ...
- Problématiques, cours, recherche etc souvent dominés par une logique d'entreprise ou une certaine vision du politique (gouvernance par le nombre, A.Soupiot)
- Besoin de reintroduire standards académiques, multidisciplinarité

Efforts Science des Données en Idf : Recherche

• PSL

Appel à projets de recherche dans le domaine de l'informatique, du calcul scientifique et des grands flots de données (« Big Data ») : l'objectif est de structurer au sein de PSL une communauté de recherche dans ces domaines en faisant émerger des projets innovants qui seront réalisés dans le cadre d'une collaboration entre au moins deux équipes hébergées par PSL.

Budget attribué pour chaque projet : entre 50000 et 100000 euros.

Ce nouvel appel à projets s'adresse à l'ensemble des chercheurs et enseignants-chercheurs travaillant dans différents laboratoires qui ont pour tutelle principale l'un des membres de PSL (hors EPST). La date limite au 1er novembre 2015.

Tags: APPEL À PROJET



Accueil Actualités Projets Postes Connexion

Le Center for Data Science à Paris-Saclay: un projet pour encourager les collaborations et les synergies interdisciplinaires

Actualités

Open Software Initiative (OSI) day - Oct. 26

CDS Rapid Analytics and Model Prototyping + Training Sprints

Objectives: solving real problems and hands-on data science training

CDS Rapid Analytics and Model Prototyping + Training Sprints

Objectives: solving real problems and hands-on data science training

Le Center for Data Science de Paris-Saclay est un projet "LODC" initié par l'Université Paris-Saclay.

Science des données

Le thème de la science des données est la conception de méthodes automatisées pour analyser les données massives et complexes afin d'en extraire des informations utiles. Ce type de projet exige donc une expertise dans un vaste éventail de domaines scientifiques allant de la recherche sur les méthodes (statistiques, traitement du signal, apprentissage automatique, fouille de données, simulation de données) à la maîtrise du domaine scientifique d'où proviennent les données en passant par l'ingénierie et maintenance des logiciels.

Les objectifs du Center for Data Science (CDS)

L'objectif de ce centre est d'établir une région dans laquelle des chercheurs et des ingénieurs pourront échanger, initier et mener des projets interdisciplinaires. Ils pourront également partager leur expérience sur des projets antérieurs portant sur la science des données. Pour favoriser la synergie entre les analystes et les producteurs de données, le CDS se propose de fournir des ressources pour la mise en place des collaborations et permettre ainsi d'atténuer le risque non négligeable pris par les chercheurs souhaitant s'engager dans des projets scientifiques interdisciplinaires. L'utilisation de formes non conventionnelles comme des données de trajectoires ou de défis de données, «data challenge», sera vivement encouragée. Le CDS soutient ainsi l'initiative de l'Université Paris-Saclay l'un des acteurs majeurs des data sciences, tant sur le plan national qu'international.

La science des données dans les sciences humaines, sciences de la vie et sciences de l'ingénieur

Plus de 250 chercheurs permanents répartis dans 35 laboratoires participent au projet. Du côté mathématique et informatique, les thèmes de recherche principaux sont

- statistique,
- apprentissage automatique,
- traitement du signal,
- visualisation des données,
- bases de données.

Dans le même temps, nous nous concentrons sur des données provenant de

- Biologie et médecine,
- Astrophysique, cosmologie et astrophysique,
- Neurosciences,
- Climat,
- Traitement des langues naturelles, de textes et de musique,
- Physique des particules et des astroparticules,
- Environnement, sciences atmosphériques et océanologie,
- Économie, finance et assurance,
- Sciences sociales et éthique,
- Sciences de l'ingénieur.

Data science hits:

SU 288KE

PSL 177K

PS 97K

USPC 32,5K

(cette réunion est première en hits USPC)

Efforts Science des Données en IdF : Masters

DataScience and BigData are key enabling technologies... Data science is based on a set of techniques and theories from many fields within the broad areas of mathematics, statistics, information science, and computer science, including machine learning, signal and image processing, statistical learning, data mining, but also database, data engineering, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming and high performance computing. Of particular relevance are algorithms and methods that can be scaled to Big Data are of particular interest in data science. Data science utilizes data preparation, statistics, predictive modeling and machine learning to investigate problems. Applications are numerous: marketing, fraud detection, risk management, marketing analytics, public policy, etc. Data science has now a significant impact both in academic and applied research; it has recently proposed disruptive ideas in man-machine interface (machine translation, speech recognition), web-related technologies (search engines, ad-management, recommender systems), digital economy (marketing, fraud detection, churning), but also health care, social sciences and the humanities. It heavily influences economics, business and finance.

Schedule

Regular Week P2

- Monday, 09h00 – 12h00, **Advanced Big Data Analytics** (M. Vazirgiannis) at Polytechnique
- Wednesday, 09h00 – 12h00, **Optimization** (A. Gramfort, S. Gaïffas) at Telecom
- Wednesday, 14h00-17h00, **Méthodes probabilistes pour l'accès à l'information à grande échelle** (F. Yvon, O. Cappé) at Paris Sud
- Thursday, 09h30 – 12h30, **Statistique en grande dimension** (Ch. Giraud) at Paris Sud
- Friday, 09h00 – 12h00, **Advanced Machine Learning** (F. d'Alche-Buc, E. Le Pennec) at Polytechnique

PS

Accueil » M2 Modélisation et Méthodes Mathématiques en Economie et Finance » Structure programme

M2 Modélisation et Méthodes Mathématiques en Economie et Finance

- **Presentation**
- **Inscription/Application 2015-2016**
- **Calendrier/Calendar 2015-2016**
- **Structure programme**
- **Contact**

An introduction to Data Science and Big Data

Instructor Fabrice Rossi (Fabrice.Rossi@univ-paris1.fr)

Course Objective This course aims at clarifying the concepts associated to the very popular and often misused terms "Data Science" and "Big Data". In particular, it studies the main misuse of the terms in the media which consists in confusing general machine learning and data mining successes on the one hand and the massive data collection organized by a limited number of online actors.

Prerequisites A background in probability/statistics is needed to fully grasp the machine learning aspects. A background in computer science/programming is needed to fully grasp the computational aspects.

Course Content Outline:

- Big Data in the media versus Big Data for statisticians/compute scientists
- Computational consideration: the scoring example
- Small to medium data solutions: basic parallelism, small scale servers
- Big Data solutions: advanced parallelism, standard frameworks (e.g. Hadoop)
- Privacy and related concerns

SU

Notre atout: multidisciplinarité



UNIVERSITÉ PARIS 13

Étudiants
en L3 et M1
informatique ?

Devenez un expert en
Science des données
et en intelligence artificielle

- Apprentissage artificiel (Machine Learning)
- Fouille de masses de données (Massive Data Mining)
- Informatique décisionnelle (Business Intelligence)

Choisissez le Master EID²
Exploration Informatique des Données et Décisionnel

Une formation d'excellence vers les métiers de demain

5 bonnes raisons de choisir EID²

1. Le Master EID² est classé parmi les meilleurs masters de France en informatique décisionnelle.
2. La formation est dispensée par les meilleurs spécialistes : des universitaires et des industriels de renommée internationale.
3. Une insertion professionnelle avec 100% de taux d'embauche en CDI.
4. Une spécialisation de haut niveau dans un domaine de pointe en science des données.
5. L'acquisition de compétences scientifiques et techniques en science des données et en intelligence artificielle.

Débouchés : Data scientist, Data miner, Consultant en décisionnel, Concepteur d'outils logiciels spécialisés, Ingénieur R&D, Chercheur (thèse de doctorat).

Directeur de la formation : Professeur Younés Bennani

#UP13

www.univ-paris13.fr

CANDIDATEZ sur : www.galilee.univ-paris13.fr



Autres?

Réunion des plateformes numériques USPC 30 Octobre 2015

Puissance de calcul agrégée → Ca 5000 coeurs

Volonté de mutualisation, services

Publication de la feuille de route du numérique nationale imminente

- Mettre en place une stratégie concertée au niveau des sites de l'ESR, en définissant des indicateurs et en mettant en place un suivi budgétaire et des mécanismes de refacturation à implémenter dans le cadre des mutualisations.
 - Promouvoir Mésocentres.
 - Respecter la diversité des usages de l'ESR
 - Cohérence avec les Comues
 - Cohérence Comue-région
 - Cohérence avec ESFRI et H2020, International
 - Sécuriser le réseau des plateformes nationales et régionales
 - Soutenir l'embauche de personnel dédié au numérique
 - Logiciels collaboratifs
 - Cloud
-
- A suivre (nouvelle réunion des plateformes ?)

Alain Turing Institute

THE ALAN
TURING
INSTITUTE

[HOME](#)

[PARTNERS](#)

[NEWS](#)

[CONTACT](#)



The Alan Turing Institute is the UK's national institute for data science.

The Institute's mission is to: undertake data science research at the intersection of computer science, mathematics, statistics and systems engineering; provide technically informed advice to policy makers on the wider implications of algorithms; enable researchers from industry and academia to work together to undertake research with practical applications; and act as a magnet for leaders in academia and industry from around the world to engage with the UK in data science and its applications.

- ▶ [The Vision](#)
- ▶ [The Research](#)
- ▶ [Research Positions](#)
- ▶ [The Founding Partners](#)
- ▶ [Turing Milestones](#)
- ▶ [Partnerships](#)
- ▶ [Data Summits & Workshops](#)

THE FOUNDING PARTNERS



- **INFRA**

- EINFRA-12-2017: Data and Distributed Computing e-infrastructures for Open Science
- EINFRA-21-2017: Platform-driven e-infrastructure innovation
- EINFRA-22-2016: User-driven e-infrastructure innovation
- INFRAINNOV-01-2017: Fostering co-innovation for future detection and imaging technologies

- **MSCA**

- RISE-2016: Research and Innovation Staff Exchange
- COFUND-2016 Co-funding of regional, national and international

- ICT-01-2016: Smart Cyber-Physical Systems
- ICT-02-2016: Thin, Organic and Large Area Electronics
- ICT-03-2016: SSI - Smart System Integration
- ICT-04-2017: Smart Anything Everywhere Initiative
- ICT-06-2016: Cloud Computing
- ICT-11-2017: Collective Awareness Platforms for Sustainability and Social Innovation ..
- ICT-13-2016: Future Internet Experimentation - Building a European experimental Infrastructure
- ICT-14-2016-2017: Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation
- ICT-15-2016-2017: Big Data PPP: Large Scale Pilot actions in sectors best benefitting from data-driven innovation
- ICT-19-2017: Media and content convergence
- ICT-20-2017: Tools for smart digital content in the creative industries
- ICT-21-2016: Support technology transfer to the creative industries
- ICT-22-2016: Technologies for Learning and Skills
- ICT-23-2017: Interfaces for accessibility
- ICT-24-2016: Gaming and gamification
-

Call - CO-CREATION FOR GROWTH AND INCLUSION

- CO-CREATION-06-2017: Policy-development in the age of big data: data-driven policy-making, policy-modelling and policy-implementation
- CO-CREATION-07-2017: Towards a new growth strategy in Europe - Improved economic and social measurement, data and official statistics
- CO-CREATION-08-2016/2017: Better integration of evidence on the impact of research and innovation in policy making
- CO-CREATION-09-2016: A European map of knowledge production and co-creation in support of research and innovation for societal challenges

• CROSS CUTTING ACTIVITIES

• Call - Internet of Things

- IoT-01-2016: Large Scale Pilots
- IoT-02-2016: IoT Horizontal activities
- IoT-03-2017: R&I on IoT integration and platforms

• Call - SMART AND SUSTAINABLE CITIES

•

Strengthening the Science with and for Society Knowledge-Base

- SwafS-11-2017: Science education outside the classroom
- SwafS-13-2017: Integrating Society in Science and Innovation – An approach to co- creation
- SwafS-15-2016: Open Schooling and collaboration on science education
- SwafS-21-2017: Promoting integrity in the use of research results in evidence based policy: a focus on non-medical research
- SwafS-22-2017: The ethical dimensions of IT technologies: a European perspective focusing on security and human rights aspects
- SwafS-23-2017: Responsible Research and Innovation (RRI) in support of sustainability and governance, taking account of the international context



BERKELEY INSTITUTE FOR DATA SCIENCE

Founded in December 2013 as a result of a year+ long national selection process
\$37.8M over 5 years, along with University of Washington & NYU

- ▶ An accelerator for data-driven discovery
- ▶ An **agent of change** in the modern university as Data Science takes hold
- ▶ An incubator for the next generation of Data Science technology and practice

GORDON AND BETTY
MOORE
FOUNDATION



ALFRED P. SLOAN
FOUNDATION

"Bold new partnership launches to harness potential of data scientists and big data"



Faculty Lead/PI: **Saul Perlmutter**, Physics, Berkeley Center for Cosmological Physics

BIDS goals

- ▶ **Support meaningful and sustained interactions and collaborations**
between Methodology fields & Science domains to recognize what it takes to move these fields forward
- ▶ **Establish new Data Science career paths that are long-term and sustainable**
 - A generation of multi-disciplinary scientists in data-intensive science
 - A generation of data scientists focused on tool development
- ▶ **Build an ecosystem of analytical tools, teaching, & research practices**
 - Sustainable, reusable, extensible, easy to learn and to translate across research domains
 - Enables scientists to spend more time focusing on their science

Astronomical Data Deluge

Serious Challenge to Traditional Approaches & Toolkits

Large Synoptic Survey Telescope (LSST) - 2020

Light curves for 800M sources every 3 days
 10^6 supernovae/yr, 10^5 eclipsing binaries
3.2 gigapixel camera, 20 TB/night

LOFAR & SKA

150 Gps (27 Tflops) → 20 Pps (~100 Pflops)

Gaia space astrometry mission - 2014

1 billion stars observed ~70 times over 5 years
Will observe 20K supernovae

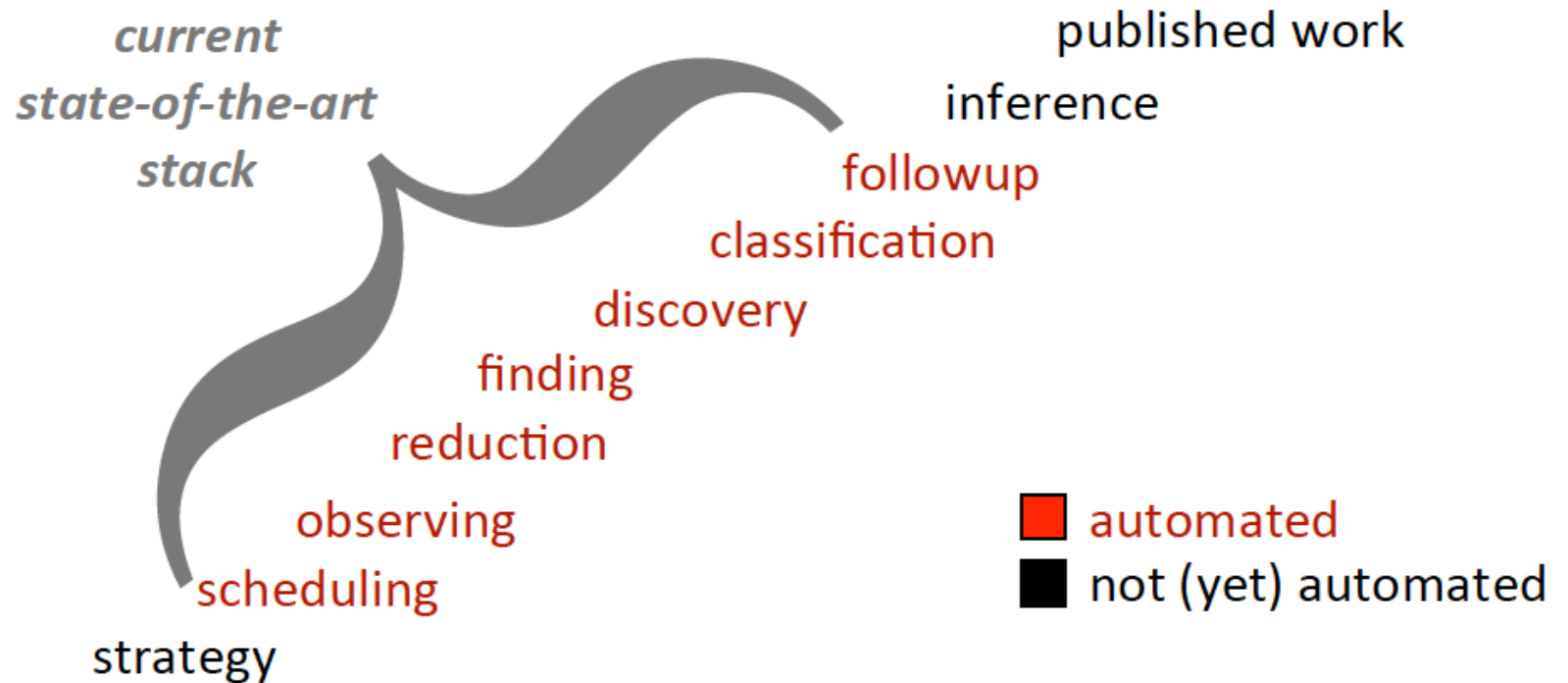
Many other astronomical surveys are already producing data:

SDSS, iPTF, CRTS, Pan-STARRS, Hipparcos, OGLE, ASAS, Kepler, LINEAR, DES etc.,

From Joshua Bloom (BIDS) presentation



Towards a Fully Automated Scientific Stack for Transients



NSF/CDI
NSF/BIGDATA

- ▶ Built & Deployed robust, real-time machine learning framework, discovering >10,000 events in > 10 TB of imaging
→ 50+ journal articles
- ▶ Built Probabilistic Event classification catalogs with innovative active learning

<http://timedomain.org>



laboratory techniques
Physics **domain
training**

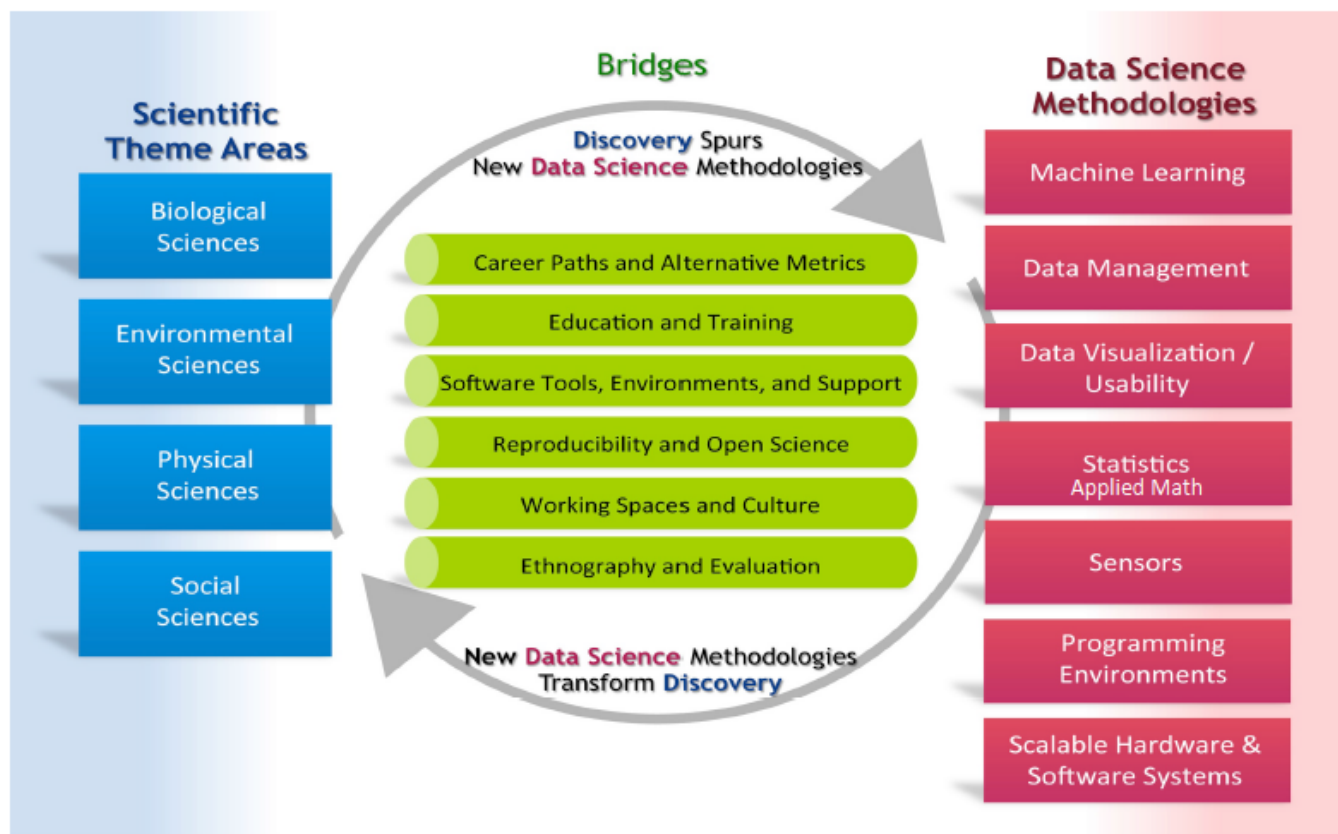
machine learning
statistics
Bayesian *MCMC*

GUI *visualization*
**advanced
computing**
database *parallel*
MapReduce

**What is the toolbox
of the modern
(data-driven) scientist?**

Peut-on construire un cursus avec plusieurs domaines de départ?

Berkeley/UW/NYU Working Groups as Bridges





4 hubs régionaux pour doper l'innovation dans le big data

📅 Publié le jeudi 19 novembre 2015 , [Partager](#)

L'administration Obama avait lancé, dès mars 2012, une campagne ambitieuse de financement de la recherche dans les domaines du big data [1]. Dotée de 200M\$, cette initiative conjointe de la Maison Blanche (OSTP), de la *National Science Foundation* (NSF) avec le soutien de 6 grandes agences fédérales, avait notamment permis de proposer des appels à projets thématiques autour des applications en santé ou en géosciences, de bâtir des programmes de sensibilisation et de formation, ou de financer à hauteur de 10M\$ le

centre de recherches AMP (*Algorithms/Machines/People*) de l'Université de Berkeley [2]. C'est dans ce laboratoire d'excellence, dirigé par Michael Jordan et Michael Franklin, qu'a notamment été mise au point la suite open source complète BDAS (Berkeley Data Analytics Stack), qui permet d'analyser les très grands volumes de données pour des applications en oncologie ou en énergie.

Dans le sillage de ce chantier national autour de l'innovation dans le big data, après avoir financé de la R&D, des infrastructures et des programmes de formation, le département CISE [3] de la NSF avait souhaité, en début d'année, structurer géographiquement la communauté des acteurs publics et privés impliqués sur ces thématiques liées aux grands volumes de données. Un appel à projets avait ainsi été lancé en mars 2015 [4] pour constituer 4 consortia régionaux, appelés *BD Hubs* (*Big Data Regional Hubs*), ayant pour vocation d'animer le tissu de toutes les parties prenantes, et de réfléchir ensemble à des problématiques big data plus spécifiques aux zones du pays concernées. Chacun de ces réseaux régionaux pouvait prétendre à un financement maximal de 1,25M\$ sur 3 ans pour composer une équipe et financer des activités d'animation, de sensibilisation, de diffusion des bonnes

- **Hub Nord-Est** (9 Etats), piloté par Columbia University, focalisé sur l'énergie, la finance, l'éducation, le climat et l'environnement
- **Hub Sud** (16 Etats), piloté par Georgia Institute of Technology et University of North Carolina, focalisé sur la santé publique, l'industrie, les matériaux, les risques côtiers
- **Hub Midwest** (12 Etats), piloté par University of Illinois at Urbana Champaign, focalisé sur l'agriculture, l'eau, les villes intelligentes
- **Hub Ouest** (13 Etats), piloté par UC San Diego, UC Berkeley et University of Washington, focalisé sur la gestion des ressources naturelles, les technologies hardware sous-jacentes, le machine learning et la médecine de précision.

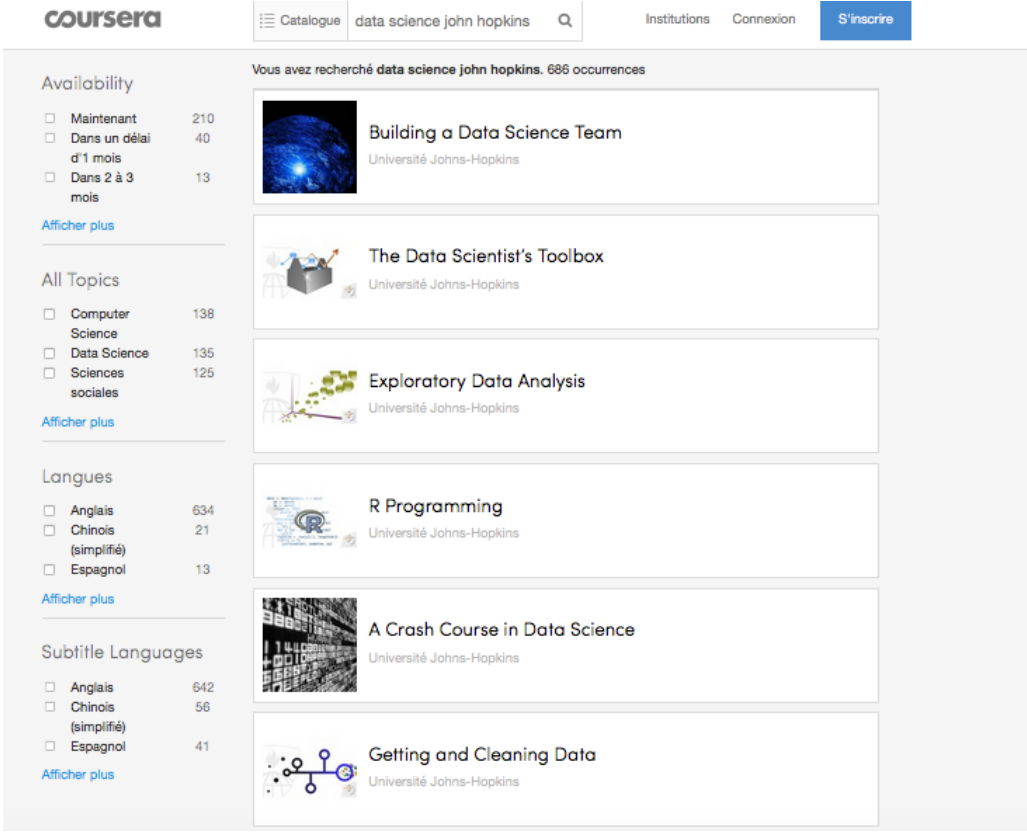
1. Introduction to Data-Science – University of Washington – Coursera Programme : Manipulation de données volumineuses, introduction à Hadoop, MapReduce et NoSQL. Modélisation statistique, Machine learning, Graph Analytics, Text-Mining, Filtres Collaboratifs. Communication des résultats, DataViz

2. MIT Online X Program : Tackling the Challenges of Big Data
Le MOOC du MIT lancé début 2014 couvre l'état de l'art des technologies et des applications Big Data en entreprise. Le programme étudie la collecte de données, le stockage et le processing de données, l'extraction et l'analyse de données non structurées, les analytics (machine learning, compression, algorithmes), la visualisation, et les nombreuses applications possibles.

3. Machine Learning – Stanford – Coursera

Supervised learning : Unsupervised learning Best practices in machine learning :

John Hopkins Data Science sous Coursera



The screenshot shows the Coursera website interface. At the top, the Coursera logo is on the left, and navigation links for 'Catalogue', 'Institutions', 'Connexion', and 'S'inscrire' are on the right. A search bar contains the text 'data science john hopkins'. Below the search bar, a message states 'Vous avez recherché data science john hopkins. 686 occurrences'. The main content area displays a list of search results, each with a thumbnail image, a title, and the institution name 'Université Johns-Hopkins'. The results include:

- Building a Data Science Team**
- The Data Scientist's Toolbox**
- Exploratory Data Analysis**
- R Programming**
- A Crash Course in Data Science**
- Getting and Cleaning Data**

On the left side of the search results, there are three filter sections:

- Availability**:
 - ☐ Maintenant 210
 - ☐ Dans un délai d'1 mois 40
 - ☐ Dans 2 à 3 mois 13
- All Topics**:
 - ☐ Computer Science 138
 - ☐ Data Science 135
 - ☐ Sciences sociales 125
- Langues**:
 - ☐ Anglais 634
 - ☐ Chinois (simplifié) 21
 - ☐ Espagnol 13

Each filter section has an 'Afficher plus' link below it. At the bottom of the filters, there is a 'Subtitle Languages' section with the following options:

- ☐ Anglais 642
- ☐ Chinois (simplifié) 56
- ☐ Espagnol 41

There is also an 'Afficher plus' link for the subtitle languages.

Sapiens et USPC



[A propos de](#) [Idex](#) [Formation](#) [Recherche](#) [International](#) [Bibliothèques](#) [Vie de campus](#) [Médias](#)

[Appels à projets](#) [Offres de formation des établissements](#) [CFDip](#) [SAPIENS](#) [PERL](#)

Formations pédagogiques

Formation à distance et MOOCs

Accompagnement individuel

Contacts SAPIENS

S'abonner à la Lettre

Trois nouveaux MOOC de USPC seront ouverts à la rentrée 2015

JEUDI, 09 JUILLET 2015 14:06



Trois nouveaux cours en ligne, qui ont bénéficié de l'accompagnement du service SAPIENS de USPC, seront ouverts au public à la rentrée 2015 : « Cities are back in town: urban sociology for a globalizing urban world » de Patrick LE GALÈS (Sciences Po), « Volcanologie physique : des phénomènes au processus » d'Edouard KAMINSKI (IPGP) et « Numérique et recherche en santé et sciences du vivant », conçu et réalisé par Celya GRUSON-DANIEL (Centre Virchow

Villermé).

[Lire la suite...](#)



[Courses](#)

[About](#)

[Partners](#)

[Sign In](#)

[Register](#)

FREE ONLINE COURSE

Gravity! From the Big Bang to Black Holes

Gravity runs the Universe. This free online course explains why, focussing on key concepts from the Big Bang to black holes.

[Join now – starts 26 Oct](#)

Un succès de USPC
70000 inscrits !

Futurelearn



MASSIVELY MULTIPLAYER ONLINE SCIENCE

A new innovative citizen science platform connecting scientific research and video games

[LEARN MORE](#)[NEWS](#)

Massively Multiplayer Online Science **connects scientific research and video games as a seamless gaming experience.** Research tasks completely integrated with game mechanics, narrative and visuals can open up a new channel between the gamer and the scientific community. Converting a small fraction of the billions of hours spent with playing video games will bring an enormous contribution to scientific research, and in the meantime will change how video games' expertise is perceived.

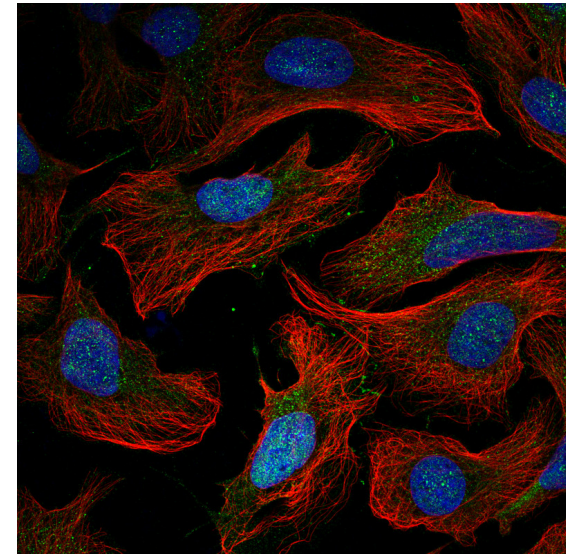
EVE ONLINE & MMOS

The first implementation is soon to arrive

[EVE ONLINE](#)[READ ABOUT THE PROJECT](#)

EVE online is the biggest science-fiction themed MMOG created by the Icelandic company, CCP. In EVE Online the players become capsuleers - spaceship pilots - who live, fight, trade, manufacture, die and get born again in the virtual EVE Universe. It is a sandbox game, which means, that all the players are playing on the same server, and everything that happens in the EVE Universe, becomes part of the history of EVE.

The first project that we chose was a classification problem of immunofluorescent images from the Human Protein Atlas. The task will be to find one or more locations in the cell where proteins are attached to it. The problem is relatively simple, though our experience with testing the feature tells us, that sometimes to find the good answer really requires attention, learning and experience. <https://www.youtube.com/watch?v=AdfFnTt2UT0>



Première liste d'actions

1. Groupe de réflexion création d'un master science des données
 1. Commencer par un MOOC pluridisciplinaire
 2. Appel pour un MMOS ?
2. Groupes de travail recherche, quel découpage ?
3. Groupes de travail propositions H2020
 1. Identifier les partenaires nationaux, Européens, internationaux
 2. Nouvelle réunion avec partenaires potentiels
4. Groupe de réflexion
 1. Logiciel de construction de communauté (de logiciels USPC à Google+)
 2. Mutualisation de plateformes numériques selon usage.
 3. Implémentation de la feuille de route nationale du numérique
 4. Valorisation