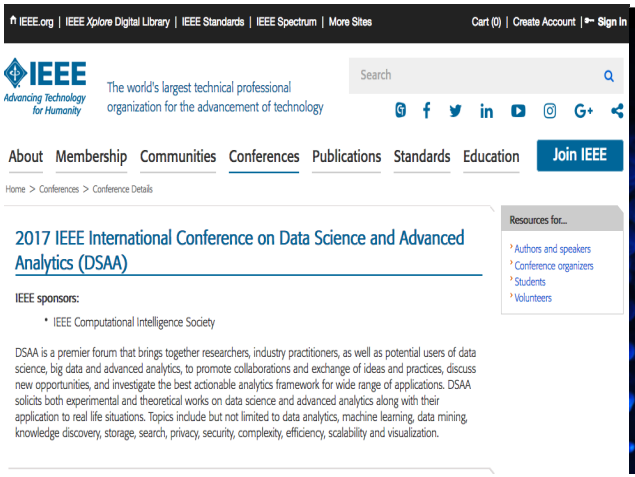
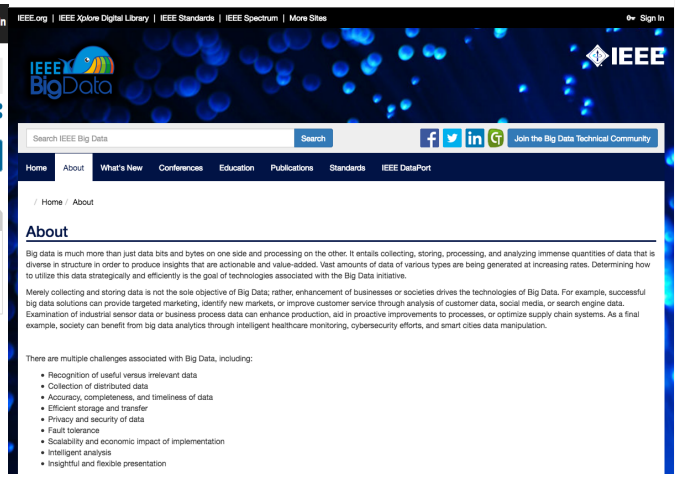
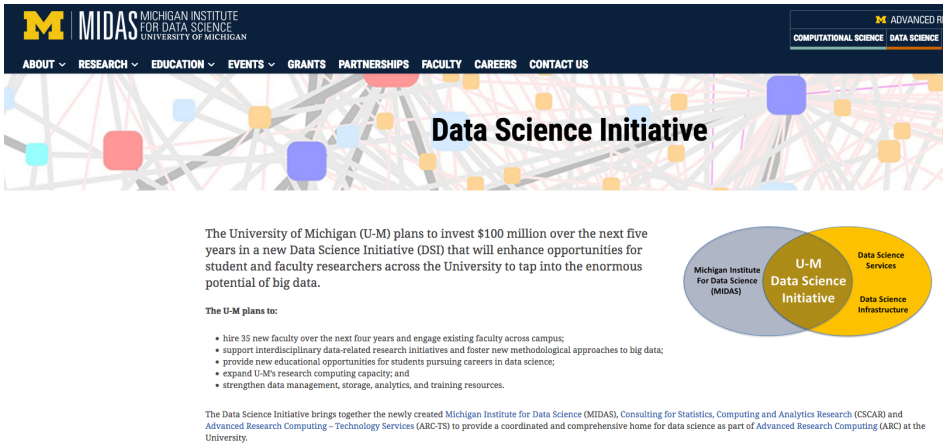
Research Summer School in Statistics & BigData Science
SBDS 2017, Caen, France

Faicel Chamroukhi

- The term “Data Science” has surged in popularity
- Data science is increasingly commonly used with “big data.”
- Data science, including Big Data has recently attracted an enormous interest from the scientific community








Le CNRS | Annuaires | Mots-Clefs CNRS | Autres sites
Mission interdisciplinarité
Centre national de la recherche scientifique
La Mission pour l'interdisciplinarité
Actualités interdisciplinaires
Les défis
Actions de site
PEPS en réseau
Plateforme réseaux
Colloques et manifestations
Espace communication
Archives
Rechercher
Sur ce site
Sur le Web du CNRS
Accueil du site > Les défis > MASTODONS
Grandes masses de données scientifiques - MASTODONS
Année de démarrage : 2012
Contexte :
Nos capacités de production, d'acquisition et d'accès aux données dépassent largement nos capacités de traitement, d'analyse et d'archivage. De nombreux domaines scientifiques (ex : cartographie du génome, observatoire virtuel en astronomie, simulation en physique et énergie), économiques (ex : e-commerce, systèmes décisionnels), ou sociétaux (ex : réseaux sociaux, bibliothèques numériques, distribution de contenus multimédia) produisent et consomment des volumes de données considérables. Le CNRS, à travers ses unités de recherche et ses grands instruments, capitalise plusieurs centaines, voire milliers, de bases de données et de corpus d'informations dont les volumes croissent de façon exponentielle.
Objectifs :
Le but est d'identifier et de soutenir des actions de recherche dont les résultats ne pourraient être obtenus sans une fertilisation croisée des disciplines et sans une synergie effective entre chercheurs. Ces actions de recherche doivent couvrir aussi bien les verrous scientifiques posés par l'utilisation à grande échelle des données (sémantique, stockage, recherche, visualisation, ...) que l'impact sociétal qu'induisent les technologies proposées (protection de la vie privée, préservation de la connaissance, ...).
Les projets sélectionnés couvrent les axes de recherche suivants :
1. Stockage, indexation et accès aux données (en particulier dans le Cloud),
2. Extraction de connaissances, apprentissage et la visualisation de grandes masses de données,
3. Qualité des données, confidentialité et sécurité des données.
4. Problèmes de propriété, de droit d'usage, droit à l'oubli.
5. Préservation/archivage des données pour les générations futures.
Colloque de restitution du défi Mastotons
Défi Mastodons - Les Big Data en recherche - AAP 2017
Grandes masses de données scientifiques - MASTODONS
AAP 2016 - La qualité des données dans les Big Data : résultats


MaDICS
Masses de Données, Informations et Connaissances en Sciences
Big Data - Data Science
Accueil
Actions
Manifestations
Réseaux
Actualités
Offres d'emploi
Calendrier
Nous contacter
Intranet
Présentation MaDICS
Se Connecter
Username
Password
Login →
Devenir membre
Mot de passe oublié
MaDICS On Twitter
Tweets by @GdR_MaDICS
Prochainement


ANR The French National Research Agency
Return to homepage
science
About ANR
Funding opportunities
Projects and Results
International cooperation
Information > News >
Information
News
Agenda
Documents
Press area
Useful Links
Big data and social sciences and humanities: A call for Transnational Research Proposals soon to be launched
DIGGING INTO DATA CHALLENGE
The French National Research Agency will launch, in partnership with 10 countries from three continents, the first international grant program of the Trans-Atlantic Platform (T-AP).
The T-AP "Digging into Data" Challenge is scheduled to be open on March. The Challenge will support researchers in examining how "big data," and the methods to use it, are changing the social sciences and humanities research landscape.
The grant program is open to projects that address any research question in social sciences and/or humanities disciplines by using new, large-scale, digital data analysis techniques. All projects must show how these techniques can lead to new theoretical insights. Proposed projects can use any data source.
Application details for the upcoming Challenge will be made available on the Digging into Data Challenge website as of March 1, 2016.
About T-AP
Trans-Atlantic Platform
Social Sciences and Humanities
The Trans-Atlantic Platform is an unprecedented collaboration between key humanities and social science funders and facilitators from South America, North America and Europe.
T-AP aims to enhance the ability of funders, research organizations and researchers to engage in transnational dialogue and collaboration. Among other activities, it works to identify common challenges and promote a culture of digital scholarship in social science and humanities research. T-AP also facilitates the formation of networks within the social sciences and humanities and helps connect them with other disciplines.
About the Digging into Data Challenge
The Digging into Data Challenge has been funding cutting-edge digital research in the social sciences and humanities since 2009. Now under the auspices of T-AP, the program will support collaborative research teams from three continents: Europe (Finland, France, Germany, the Netherlands, Portugal (to be confirmed) and the United Kingdom); North America (Canada, Mexico, the United States); and South America (Brazil and Argentina).


université PARIS-SACLAY
Paris-Saclay Center for Data Science
ABOUT
PEOPLE
PROJECTS
EVENTS
NEWS
JOBS
CONTACT
Search
Tweets by SaclayCDS
PARIS-SACLAY Center for Data Science (CDS)
Phase I : Lidex Paris-Saclay (2014 – 2016)
Phase II : IRS Initiatives de Recherche Stratégiques (2016 – 2019)
Extracting knowledge from data.
The project consists of developing methods and tools so as to be capable of analysing gigantic amounts of data and extracting useful information from them for physics, biology, medicine, chemistry, the environment and the human sciences.
This project is multidisciplinary; it requires research on analytical methodologies (statistics, processes of machine learning, extracting knowledge, viewing data), as well as on software design.
More than 250 permanent researchers in 35 laboratories participate in the CDS supporting our datascience projects and events.
News
Associate/full professor position in the area of Computer Vision 2017/03/20
Appel à projets, Paris-Saclay Center for Data Science 2017/03/14
Wikidata pour la science 2017 2017/03/14
Permanent position of Professor in Signal Processing at CentraleSupelec @ Université Paris-Saclay 2017/03/13
Permanent position (Associated Professor) in Machine Learning @ TELECOM-ParisTech 2017/03/13
Appel à projets émergents 2017 du département STIC 2017/03/01
One day Workshop and Hackathon on spatio-temporal time series

- What does Data Science mean?
- What about Statistics in the Data Science "area" ?
- There is not yet a consensus on what precisely constitutes Data Science

CONTRIBUTED ARTICLES

Data Science and Prediction

By Vasant Dhar
Communications of the ACM, Vol. 56 No. 12, Pages 64-73
10.1145/2500499
Comments (2)

VIEW AS: SHARE:

Use of the term "data science" is increasingly common, as is "big data." But what does it mean? Is there something unique about it? What skills do "data scientists" need to be productive in a world deluged by data? What are the implications for scientific inquiry? Here, I address these questions from the perspective of predictive modeling.

Back to Top

Key Insights

- Data science is the study of the generalizable extraction of knowledge from data.
- A common epistemic requirement in assessing whether new knowledge is actionable for decision making is its predictive power, not just its ability to explain the past.
- A data scientist requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions.

ASA | *Amstat News* | ASA Community | The World of Sta

AMSTATNEWS
The Membership Magazine of the American Statistical Association

HOME | ABOUT | EDITORIAL CALENDAR | PDF ARCHIVES | ADVERTISE | STATISTICIANS IN HISTORY

Home » Featured

ASA Statement on the Role of Statistics in Data Science

1 OCTOBER 2015 6,956 VIEWS 13 COMMENTS

Statement Contributors
David van Dyk, *Imperial College* (chair)
Montse Fuentes, *NCSU*
Michael I. Jordan, *UC Berkeley*
Michael Newton, *University of Wisconsin*
Bonnie K. Ray, *Pegged Software*
Duncan Temple Lang, *UC Davis*
Hadley Wickham, *RStudio*

The rise of data science, including Big Data and data analytics, has recently attracted enormous attention in the popular press for its spectacular contributions in a wide range of scholarly disciplines and commercial endeavors. These successes are largely the fruit of the innovative and entrepreneurial spirit that characterize this burgeoning field. Nonetheless, its interdisciplinary nature means that a substantial collaborative effort is needed for it to realize its full potential for productivity and innovation. While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science: (i) Database Management enables transformation, conglomeration, and organization of data resources, (ii) Statistics and Machine Learning convert data into knowledge, and (iii) Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.

- For a review, see the report of D. Donoho (2015): "50 years of Data Science"

INSTITUT DE FRANCE
Académie des sciences

recherche ...

NOUS CONNAÎTRE | VIE SCIENTIFIQUE | ENSEIGNEMENT DES SCIENCES | DIFFUSION DES CONNAISSANCES | COLLABORATIONS INTERNATIONALES | EXPERTISE ET CONSEIL

18 FÉV 2014

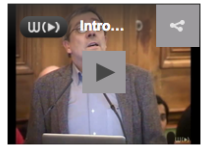
La datamasse : directions et enjeux pour les données massives

Publié dans Colloques, conférences et débats

Conférence-débat de l'Académie des sciences

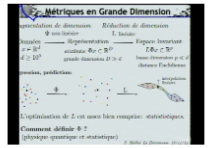
Nous vivons dans une "société de l'information" dont les avancées scientifiques et techniques rapides, associées au développement d'usages nouveaux, conduisent à produire des quantités toujours plus gigantesques de données numériques. Cette situation d'abondance ouvre des perspectives nouvelles tant dans les sciences exactes que dans les sciences humaines. L'utilisation de cette "datamasse" (*Big Data* en anglais) pose des défis considérables : Comment stocker de telles quantités de données, les manipuler, les analyser, les trier,... les valoriser ? Comment concilier leur omniprésence et le respect de la vie privée ? Comment faire qu'elles bénéficient à tous ? Ce sont quelques-uns de ces aspects qui seront mis en avant dans cette rencontre, afin d'en mieux comprendre les possibilités et les limitations, pour en mieux maîtriser les développements.

Introduction
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences et Patrick Flandrin, directeur de recherche CNRS, École normale supérieure de Lyon, membre de l'Académie des sciences

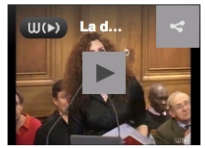


À la découverte des connaissances massives de la Toile
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences

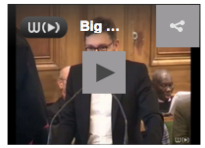
Des mathématiques pour l'analyse de données massives
Stéphane Mallat, professeur à l'École normale supérieure, Paris



La découverte du cerveau grâce à l'exploration de données massives
Anastasia Ailamaki, professeure à l'École polytechnique fédérale de Lausanne



Big Data et Relation Client : quel impact sur les industries et activités de services traditionnelles ?
François Bourdoncle, co-fondateur et CTO d'Exalead, filiale de Dassault Systèmes



Discussion générale et conclusion

Vidéos réalisées par la cellule Webcast CC-IN2P3 du CNRS

- There is not yet a consensus on what precisely constitutes Data Science, but
- Data Science can be seen (defined ?) as[a]:
  - the study of the generalizable extraction of knowledge from data.
  - requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization

[a] Vasant Dhar (2013): Communications of the ACM, Vol. 56 No. 12: 64-73

- Data Science clearly has an interdisciplinary nature and requires substantial collaborative effort
- Databases, statistics and machine learning, and distributed systems are emerging as foundational to data science

(i) Databases: organization of data resources,

(ii) Statistics and Machine Learning: convert data into knowledge,

(iii) Distributed and Parallel Systems: computational infrastructure

Statistics play a central role in data science

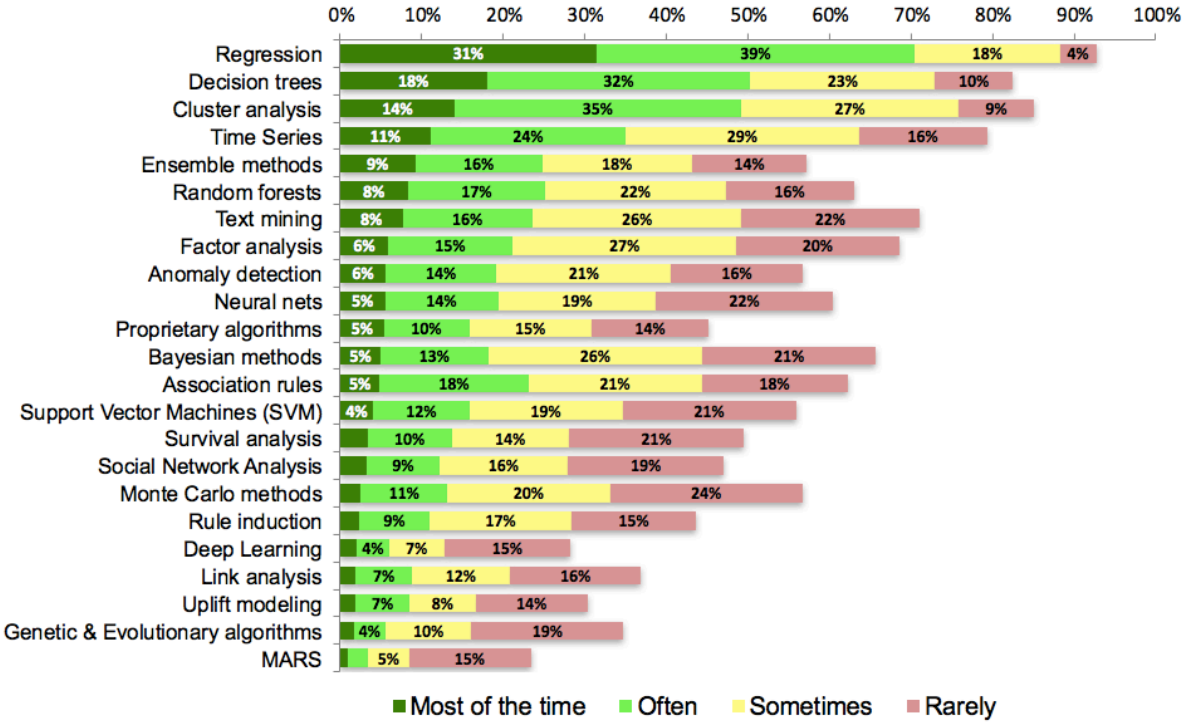
- Allow to quantify the randmoness component in the data
- A well-established background to deal with uncertainty (probabilistic framework) and to establish generizable methods for prediction and estimation
- allow soft decision: e.g. confidence interval in regression and posterior probabilities in classification
- help for understanding the underlying generative process

Data science models/algorithms

New problems (big data, etc) but ... classical methods ?

Our Core Algorithms Remain the Same

- Regression, decision trees, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent since the first Data Miner Survey in 2007.



Question: What algorithms / analytic methods do you TYPICALLY use? (Select all that apply)

2016 Rexer Analytics 4

SBDS 2017 : Research Summer School in Statistics & BigData Science (SBDS)
7-9 June @ Caen

Home
Lecturers
Schedule
Talks & Demos
Registration
Practical Information
Sponsors
Contact

Christophe Ambroise

- Professor, Evry University, France
- **Talk:** Statistical learning of stochastic latent block models for networks inference

Peter Tino

- Pofessor, University of Birmingham, UK
- **Talk:** Probabilistic Modelling in Machine Learning

Romain Hérault

- Associate Pofessor, National Institute of Applied Sciences of Rouen, France
- **Talk:** Deep Learning

Jalal Fadili

- Professor, Ensicaen & Insitut Universitaire de France (IUF), France
- **Talk:** Sparse representation of high dimensional signals and images

Hien Nguyen

- Australian Research Council DECRA Research Fellow, La Trobe University, Australia
- **Talk:** An introduction to MM algorithms for the machine learning and statistical estimation

Abstract: MM (majorization-minimization) algorithms are an increasingly popular tool for solving optimization problems in machine learning and statistical estimation. This lecture introduces the MM algorithm framework in general and via three commonly considered example applications: Gaussian mixture models, multinomial logistic regressions, and support vector machines. Specific algorithms for these three examples are derived and numerical demonstrations are presented. Theoretical and practical aspects of MM algorithm design are discussed.

Mustapha Lebbah

- Associate Professor, Paris 13 University
- **Talk:** Scalable machine learning and distributed systems

Faicel Chamroukhi

- Professor, Caen University, France
- **Talk:** Unsupervised learning of latent variable models from high-dimensional data

Sponsors

Modal: MOdels for Data Analysis and Learning

Thank you for your attention!