Research Summer School on Statistics for Data Science S4D 2018, Caen, France

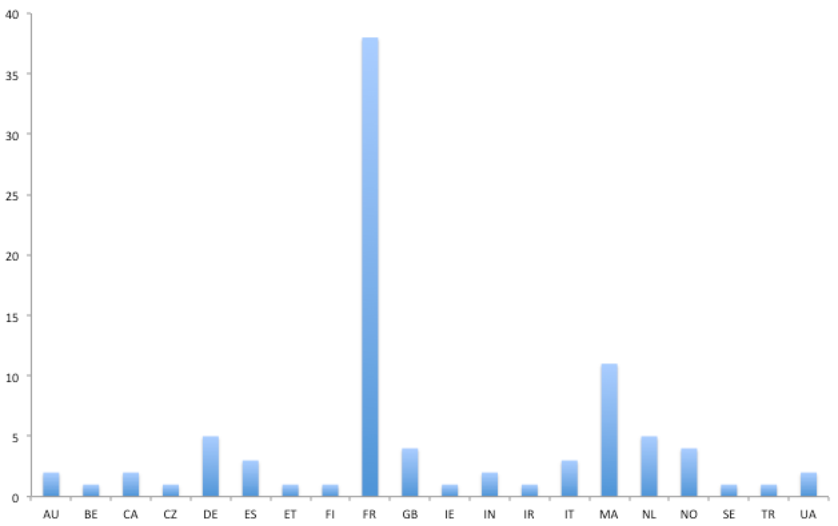
Faicel Chamroukhi
https://chamroukhi.com

S4D key insights

- Talks covering both tutorial and advanced aspects at the interface of Statistics, Machine Learning and Optimization

  $\hookrightarrow$ the main data science fields
- Theoretical foundations and algorithmic aspects, as well as typical case studies in complex and large-scale scenarios
- We enjoy good food and nice visits to Normandie sites :)

86 participants from 20 countries!



Many thanks to our sponsors!

RÉGION NORMANDIE

LMNO

UNICAEN

Normandie Université

SFdS

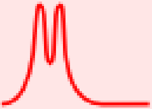
Caen la mer
NORMANDIE
COMMUNAUTÉ URBAINE

Normandie - Mathématiques

Normastic

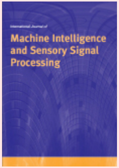
cnrs dépasser les frontières

MaDICS



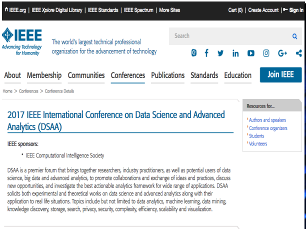
LMRS

SUMMER
SCHOOLS
IN
EUROPE
.EU



International Journal of
Machine Intelligence
and Sensory Signal
Processing

- 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

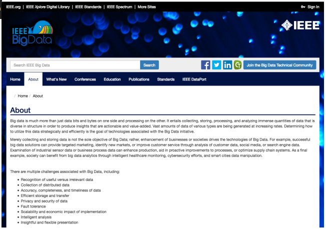

IEEE.org | IEEE Xplore Digital Library | IEEE Standards | IEEE Spectrum | More Sites
Cart (0) | Create Account | Sign In
IEEE Advancing Technology for Humanity
The world's largest technical professional organization for the advancement of technology
Search
About
Membership
Communities
Conferences
Publications
Standards
Education
Join IEEE
Home > Conferences > Conference Details
Resources for...
Authors and speakers
Conference organizers
Students
Volunteers

2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)

IEEE sponsors:

- IEEE Computational Intelligence Society

DSAA is a premier forum that brings together researchers, industry practitioners, as well as potential users of data science, big data and advanced analytics, to promote collaborations and exchange of ideas and practices, discuss new opportunities, and investigate the best actionable analytics framework for wide range of applications. DSAA solicits both experimental and theoretical works on data science and advanced analytics along with their application to real life situations. Topics include but not limited to data analytics, machine learning, data mining, knowledge discovery, storage, search, privacy, security, complexity, efficiency, scalability and visualization.



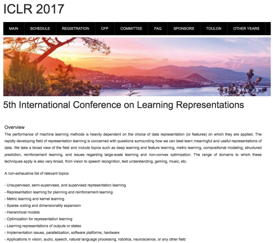
About

Big data is much more than just data bits and bytes on one side and processing on the other. It entails collecting, storing, processing, and analyzing immense quantities of data that is diverse in structure in order to produce insights that are actionable and value-added. Vast amounts of data of various types are being generated at increasing rates. Determining how to utilize this data strategically and efficiently is the goal of technologies associated with the Big Data initiative.

Merely collecting and storing data is not the sole objective of Big Data; rather, enhancement of businesses or societies drives the technologies of Big Data. For example, successful big data solutions can provide targeted marketing, identify new markets, or improve customer service through analysis of customer data, social media, or search engine data. Examination of industrial sensor data or business process data can enhance production, aid in proactive improvements to processes, or optimize supply chain systems. As a final example, society can benefit from big data analytics through intelligent healthcare monitoring, cybersecurity efforts, and smart cities data manipulation.

There are multiple challenges associated with Big Data, including:

- Recognition of useful versus irrelevant data
- Collection of distributed data
- Accuracy, completeness, and timeliness of data
- Efficient storage and transfer
- Privacy and security of data
- Fault tolerance
- Scalability and economic impact of implementation
- Intelligent analysis
- Insightful and flexible presentation



ICLR 2017

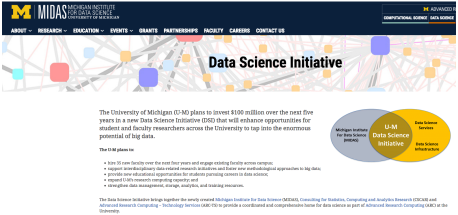
5th International Conference on Learning Representations

Overview

The performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied. The rapidly developing field of representation learning is concerned with questions surrounding how we can best learn meaningful and useful representations of data. We take a broad view of the field and include topics such as deep learning and feature learning, metric learning, compositional modeling, structured prediction, reinforcement learning, and issues regarding large-scale learning and non-convex optimization. The range of domains to which these techniques apply is also very broad, from vision to speech recognition, text understanding, gaming, music, etc.

A non-exhaustive list of relevant topics:

- Unsupervised, semi-supervised, and supervised representation learning
- Representation learning for planning and reinforcement learning
- Metric learning and kernel learning
- Sparse coding and dimensionality expansion
- Hierarchical models
- Optimization for representation learning
- Learning representations of outputs or states
- Implementation issues, parallelization, software platforms, hardware
- Applications in vision, audio, speech, natural language processing, robotics, neuroscience, or any other field


M | MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN
M ADVANCED R
COMPUTATIONAL SCIENCE
DATA SCIENCE
ABOUT
RESEARCH
EDUCATION
EVENTS
GRANTS
PARTNERSHIPS
FACULTY
CAREERS
CONTACT US
Michigan Institute For Data Science (MIDAS)
U-M Data Science Initiative
Data Science Services
Data Science Infrastructure

Data Science Initiative

The University of Michigan (U-M) plans to invest $100 million over the next five years in a new Data Science Initiative (DSI) that will enhance opportunities for student and faculty researchers across the University to tap into the enormous potential of big data.

The U-M plans to:

- hire 35 new faculty over the next four years and engage existing faculty across campus;
- support interdisciplinary data-related research initiatives and foster new methodological approaches to big data;
- provide new educational opportunities for students pursuing careers in data science;
- expand U-M's research computing capacity; and
- strengthen data management, storage, analytics, and training resources.

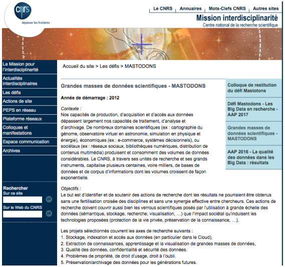
The Data Science Initiative brings together the newly created Michigan Institute for Data Science (MIDAS), Consulting for Statistics, Computing and Analytics Research (CSCAR) and Advanced Research Computing – Technology Services (ARC-TS) to provide a coordinated and comprehensive home for data science as part of Advanced Research Computing (ARC) at the University.

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE



Le CNRS | Annuaires | Mots-Clefs CNRS | Autres sites

Mission interdisciplinarité

Centre national de la recherche scientifique

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, …).

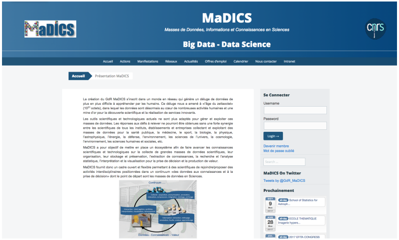
Les projets sélectionnés couvrent les axes de recherche suivants :
1. Stockage, indexation et accès aux données (en particulier dans le Cloud),
2. Extraction de connaissances, apprentissage et la visualisation de grandes masses de données,
3. Qualité des données, confidentialité et sécurité des données.
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

La création du GdR MaDICS s'inscrit dans un monde en réseau qui génère un déluge de données de plus en plus difficile à appréhender par les humains. Ce déluge nous a amené à «l'âge du zettaoctet» (10^{21} octets), dans lequel les données sont désormais au cœur de nombreuses activités humaines et une mine d'or pour la découverte scientifique et la réalisation de services innovants.

Les outils scientifiques et technologiques actuels ne sont plus adaptés pour gérer et exploiter ces masses de données. Les réponses aux défis à relever ne pourront être obtenues sans une forte synergie entre les scientifiques de tous les instituts, établissements et entreprises collectant et exploitant des masses de données pour la santé publique, la médecine, le sport, la biologie, la physique, l'astrophysique, l'énergie, la défense, l'environnement, les sciences de l'univers, la cosmologie, l'environnement, les sciences humaines et sociales, etc.

MaDICS a pour objectif de mettre en place un écosystème afin de faire avancer les connaissances scientifiques et technologiques sur la collecte de grandes masses de données scientifiques, leur organisation, leur stockage et préservation, l'extraction de connaissances, la recherche et l'analyse statistique, l'interprétation et la visualisation pour la prise de décision et la production de valeur.

MaDICS fournit donc un cadre ouvert et flexible permettant à des scientifiques de rejoindre/proposer des activités interdisciplinaires positionnées dans un continuum «des données aux connaissances et à la prise de décision» dont le point de départ sont les masses de données en Sciences.

Se Connecter

Username

Password

Login →

Devenir membre
Mot de passe oublié

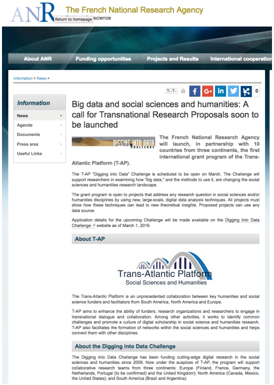
Search …

MaDICS On Twitter

Tweets by @GdR_MaDICS

Prochainement

OCT 9 Mon 2017	all-day School of Statistics for Astroph…
AUG 28 Mon 2017	all-day ECOLE THEMATIQUE Imagerie hypers…
JUL	all-day 2017 EFITA CONGRESS



Information > News >

Big data and social sciences and humanities: A call for Transnational Research Proposals soon to be launched

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.

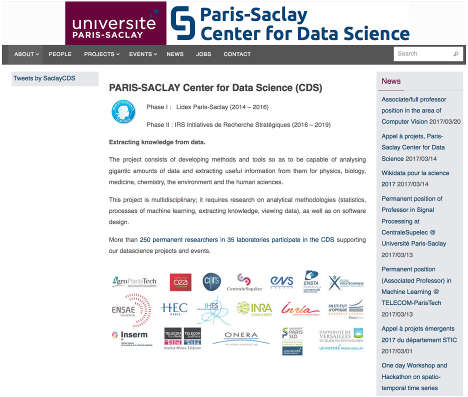
About T-AP

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).



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

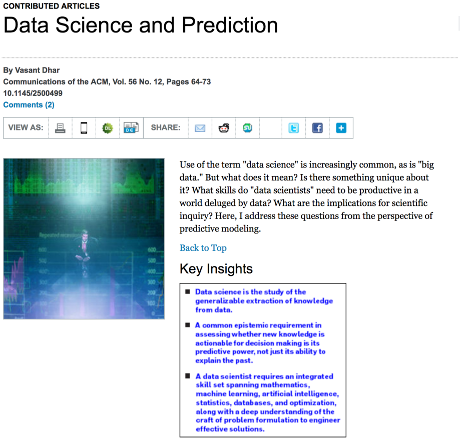
Permanent position of Professor in Signal Processing at CentraleSupelec @ Université Paris-Saclay 2017/03/13

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

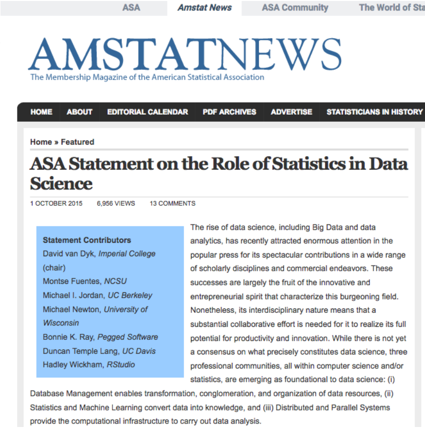
By Vasant Dhar
Communications of the ACM, Vol. 56 No. 12, Pages 64-73
10.1145/2500499
Comments (2)

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.



AMSTATNEWS
The Membership Magazine of the American Statistical Association

Home » Featured

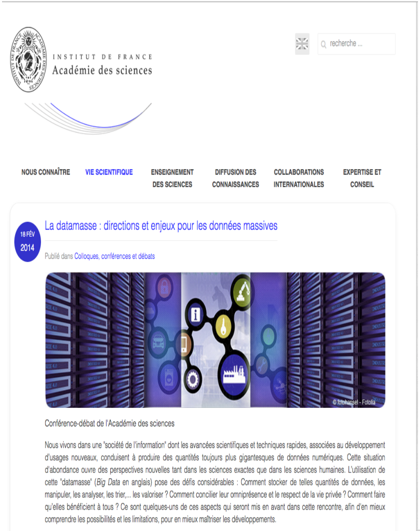
ASA Statement on the Role of Statistics in Data Science

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”



18 FÉV 2014

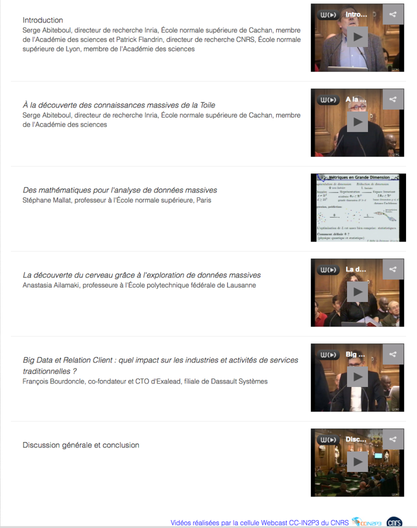
La datamasse : directions et enjeux pour les données massives

Publié dans Colloques, conférences et débats

© fotohansel - Fotolia

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,

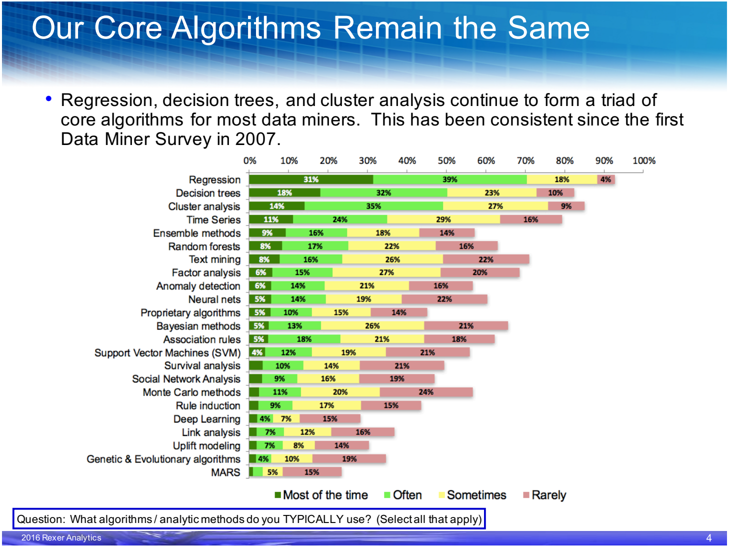
(iii) Distributed and Parallel Systems: computational infrastructure

Statistics play a central role in data science

- Allow to quantify the randomness 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 ?



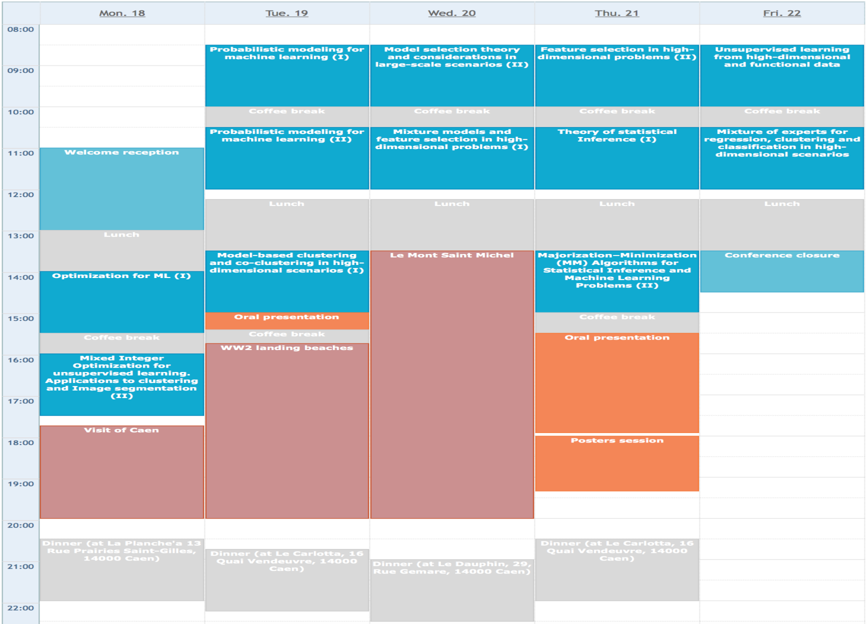
Our Core Algorithms Remain the Same

- 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

S4D 2018: programme



	Mon. 18	Tue. 19	Wed. 20	Thu. 21	Fri. 22
08:00		Probabilistic modeling for machine learning (I)	Model selection theory and considerations in large-scale scenarios (II)	Feature selection in high-dimensional problems (II)	Unsupervised learning from high-dimensional and functional data
09:00					
10:00		Coffee break Probabilistic modeling for machine learning (II)	Coffee break Mixture models and feature selection in high-dimensional problems (I)	Coffee break Theory of statistical Inference (I)	Coffee break Mixture of experts for regression, clustering and classification in high-dimensional scenarios
11:00	Welcome reception				
12:00		Lunch	Lunch	Lunch	Lunch
13:00	Lunch	Model-based clustering and co-clustering in high-dimensional scenarios (I)	Le Mont Saint Michel	Majorization–Minimization (MM) Algorithms for Statistical Inference and Machine Learning Problems (II)	Conference closure
14:00	Optimization for ML (I)				
15:00	Coffee break	Oral presentation Coffee break WW2 landing beaches		Coffee break Oral presentation	
16:00	Mixed Integer Optimization for unsupervised learning. Applications to clustering and Image segmentation (II)				
17:00	Visit of Caen				
18:00				Posters session	
19:00					
20:00	Dinner (at La Planche'a 13 Rue Prairies Saint-Gilles, 14000 Caen)	Dinner (at Le Carlotta, 16 Quai Vendeuvre, 14000 Caen)		Dinner (at Le Carlotta, 16 Quai Vendeuvre, 14000 Caen)	
21:00			Dinner (at Le Dauphin, 29, Rue Gemare, 14000 Caen)		
22:00					

Many thanks for your participation!
Enjoy the courses!
Looking forward to seeing you next year :)