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.
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.

For a review, see the report of D. Donoho (2015): “50 years of Data Science”
La datamasse : directions et enjeux pour les données massives

Conférence-clé de l’Académie des sciences

Nous vivons dans une "socité 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 limites, pour en mieux maîtriser les développements.
There is not yet a consensus on what precisely constitutes Data Science, but Data Science can be seen (defined ?) as:

▶ 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


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

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

Peter Tino
- Professor, University of Birmingham, UK
- Talk: Probabilistic Modelling in Machine Learning

Romain Héralt
- Associate Professor, National Institute of Applied Sciences of Rouen, France
- Talk: Deep Learning

Jalal Fadili
- Professor, Enscaen & Instut 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
Thank you for your attention!