INAUGURAL CONFERENCE

BIDSA - the Bocconi Institute for Data Science and Analytics

November 18, 2016 at 2:30pm
Università Bocconi, Via Sarfatti 25
Aula Manfredini

2:30pm Welcome Remarks
Gianmario Verona, Rector, Unversità Bocconi

2:50pm On Computational Thinking, Inferential Thinking and Data Science
Michael I. Jordan, University of California, Berkeley

The rapid growth in the size and scope of datasets in science and technology has created a need for novel foundational perspectives on data analysis that blend the inferential and computational sciences. That classical perspectives from these fields are not adequate to address emerging problems in "Big Data" is apparent from their sharply divergent nature at an elementary level---in computer science, the growth of the number of data points is a source of "complexity" that must be tamed via algorithms or hardware, whereas in statistics, the growth of the number of data points is a source of "simplicity" in that inferences are generally stronger and asymptotic results can be invoked. On a formal level, the gap is made evident by the lack of a role for computational concepts such as "runtime" in core statistical theory and the lack of a role for statistical concepts such as "risk" in core computational theory. I present several research vignettes aimed at bridging computation and statistics, including the problem of inference under privacy and communication constraints, and methods for trading off the speed and accuracy of inference.

3:30pm Data science and the curse of phase transitions
Marc Mézard, École normale supérieure – PSL Research University, Paris

Extracting information, and more generally extracting knowledge from large datasets is arguably one of the main frontiers of modern science, common to a broad variety of disciplines. Bayesian approaches to machine learning and signal processing provide a conceptual framework in which information bits interact through constraints (due to prior knowledge or to measurements). Statistical physics has helped to develop new approaches and very powerful algorithms in this context, where collective phenomena, like phase transitions and the occurrence of glassy phases, play a major role. This talk will review some of the main developments in this field, illustrated by specific examples like compressed sensing.

4:10pm Coffee Break

4:40pm **Large data analysis of genetic and imaging data**
Aldo Rustichini, University of Minnesota

In recent years neuroeconomic analysis of economic and strategic behavior has provided insights into the biological pathways affecting this behavior, and insights into the individual differences (for classical features in economic analysis, such as attitude to risk and time discounting, but also of Personality Traits and Intelligence). Genetic analysis mostly based on Genome Wide Association Studies (GWAS) has provided useful quantitative estimates of the association; the identification of significant Single Nucleotide Polymorphisms (SNP's) however is limited to purely correlational results.

The task of identifying pathways is made extremely difficult by the highly polygenic nature of the phenotypes of interest for economic analysis, and by the high dimensions of both imaging and genetic data. On the other hand, a correct understanding of the biological pathways to behavior of SNP's identified with GWAS to behavior is essential for economists when they have to suggest policies. An integration of the neural data (in first place, structural and functional imaging data) and genetic data is essential for future progress. The fundamental difficulty is that the polygenic nature of the phenotypes makes the candidate gene approach particularly inadequate and misleading. The method of Gene Set Enrichment Analysis (GSEA) provides a promising alternative, and has already been successfully used. We will outline a strategy based on the integration of hierarchical Bayesian models with ideas form GSEA adapted to Bayesian methodology.

5:20pm **Unreasonable Effectiveness of Learning Artificial Neural Networks**
Riccardo Zecchina, Politecnico di Torino

Deep networks are some of the most widely used tools in data science. Learning is in principle a hard problem in these systems, but in practice heuristic algorithms often find solutions with good generalization properties. We propose an explanation of this good performance in terms of a novel large-deviation measure: we show that there are regions of the optimization landscape which are both robust and accessible, and that their existence is crucial to achieve good performance on a class of particularly difficult learning problems. Building on these results, we introduce basic algorithmic schemes which improve existing optimization algorithms and provide a framework for further research on efficient learning for huge data sets and for novel computational technologies.

6:00pm Cocktail Reception