

Preface

The concept “Big data” is not only related to storage of and access to data. Analytics play a major role in making sense of that data and exploiting its value. Big data analytics is about examining vast volumes of data in order to discover hidden patterns or even existing correlations. With available technology, it is feasible to analyze data and get crucial answers from it, almost in real time.

A lot of research has been made by Google on this field with impressive results. It is worth mentioning “Google Cloud Machine Learning”, a managed platform enabling easy construction and development of machine learning models, capable to perform on any type of data, of any size. The neural network field has historically focused on algorithms that learn in an online, incremental mode without requiring in-memory access to huge amounts of data. This type of learning is not only ideal for streaming data (as in the Industrial Internet or the Internet of Things) but could also be used on stored big data. Neural network technologies thus can become significant components of big data analytics platforms.

One of the major challenges of our era is learning from big data, something that requires the development of novel algorithmic approaches. Most of the available machine learning algorithms, find it hard to scale up to big data. Moreover there are serious challenges in the problems of high-dimensionality, velocity and variety. Thus the aim of this conference is to promote new advances and research directions in efficient and innovative algorithmic approaches to analyzing big data (e.g. deep networks, nature-inspired and brain-inspired algorithms), implementations on different computing platforms (e.g. neuromorphic, GPUs, clouds, clusters) and applications of Big Data Analytics to solve real-world problems (e.g. weather prediction, transportation, energy management).

The 2nd INNS Big Data 2016 conference, founded by the International Neural Network Society (INNS), was organized by Aristotle University of Thessaloniki Greece and Democritus University of Thrace Greece, following the inaugural event held in San Francisco in 2015. This conference aims to initiate the collaborative adventure with big data and other learning technologies.

All submitted papers in the 2nd INNS Big Data conference have passed through a peer review process by at least 2 independent academic referees. Where needed a

third and a fourth referee was consulted to resolve any potential conflicts. Out of 50 submissions, totally 34 submissions have been accepted for oral presentation; 27 papers have been accepted as full ones (54 %) and 7 as short ones. Authors come from 19 different countries around the globe, namely: Brazil, China, Cyprus, Denmark, France, Germany, Greece, India, Italy, Japan, Malaysia, Portugal, Saudi Arabia, Spain, Tunisia, Turkey, USA, Ukraine and UK. Proceedings are published by Springer in the “Advances in Intelligent Systems and Computing Series”.

The program of the 2nd INNS Big Data Conference includes 5 plenary talks by distinguished keynote speakers, plus 2 tutorials.

David Bholat is Senior Analyst with the Bank of England. In particular, Dr. Bholat leads a team of ten data scientists and researchers in Advanced Analytics, a Big Data division in the Bank of England which he helped to establish in 2014. The division is recognised as a leader among central banks in the area of Big Data, as noted in a recent MIT Sloan Review article profiling the division. A former Fulbright fellow, Dr. Bholat graduated from Georgetown University’s School of Foreign Service with highest honours. He subsequently studied at the London School of Economics, the University of Chicago and London Business School. Publications in 2016 include Modelling metadata in central banks; Non-performing loans: regulatory and accounting treatments of assets; Peer-to-peer lending and financial innovation in the United Kingdom; and Accounting in central banks. Other previous publications relevant to the conference include Text mining for central banks; Big data and central banks; and The future of central bank data.

The title of the talk by Dr. Bholat is “Big Data at the Bank of England”. This talk will discuss the Bank of England’s recent forays into Big Data and other unconventional data sources. Particular attention will be given to the practicalities of embedding data analytics in a business context. Examples of the Bank’s use of Big Data will be given, including the analysis of derivatives and mortgage data, Internet searches, and social media.

Francesco Bonchi is Research Leader with the ISI Foundation of Turin and Scientific Director of the Technological Center of Catalunya at Barcelona. Before he was Director of Research at Yahoo Labs in Barcelona, Spain, where he was leading the Web Mining Research group. His recent research interests include mining query-logs, social networks, and social media, as well as the privacy issues related to mining these kinds of sensible data. In the past he has been interested in data mining query languages, constrained pattern mining, mining spatiotemporal and mobility data, and privacy preserving data mining. He will be PC Chair of the 16th IEEE International Conference on Data Mining (ICDM 2016) to be held in Barcelona in December 2016. He is member of the ECML/PKDD Steering Committee, Associate Editor of the IEEE Transactions on Big Data (TBD), the IEEE Transactions on Knowledge & Data Engineering (TKDE), the ACM Transactions on Intelligent Systems and Technology (TIST), Knowledge and Information Systems (KAIS), and member of the Editorial Board of Data Mining & Knowledge Discovery (DMKD). He has been Program Co-chair of the ECML/PKDD’2010. He is co-editor of the

book “Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques” published by Chapman & Hall/CRC Press. He earned his Ph.D. in Computer Science from the University of Pisa in December 2003.

The title of the talk by Dr. Bonchi is “On information propagation, social influence, and communities”. With the success of online social networks and microblogging platforms such as Facebook, Tumblr, and Twitter, the phenomenon of influence-driven propagations, has recently attracted the interest of computer scientists, sociologists, information technologists, and marketing specialists. In this talk we will take a data mining perspective, discussing what (and how) can be learned from a social network and a database of traces of past propagations over the social network. Starting from one of the key problems in this area, i.e. the identification of influential users, we will provide a brief overview of our recent contributions in this area. We will expose the connection between the phenomenon of information propagation and the existence of communities in social network, and we will go deeper in this new research topic arising at the overlap of information propagation analysis and community detection.

Steve Furber, CBE FRS FREng, is ICL Professor of Computer Engineering in the School of Computer Science at the University of Manchester, UK. After completing a BA in Mathematics and a PhD in Aerodynamics at the University of Cambridge, UK, he spent the 1980s at Acorn Computers, where he was a principal designer of the BBC Microcomputer and the ARM 32-bit RISC microprocessor. Over 75 billion variants of the ARM processor have since been manufactured, powering much of the world’s mobile and embedded computing. He moved to the ICL Chair at Manchester in 1990 where he leads research into asynchronous and low-power systems and, more recently, neural systems engineering, where the SpiNNaker project is delivering a computer incorporating a million ARM processors optimised for brain modelling applications.

The title of the talk by Professor Furber is “The SpiNNaker Project”. The SpiNNaker (Spiking Neural Network Architecture) project aims to produce a massively-parallel computer capable of modelling large-scale neural networks in biological real time. The machine has been 15 years in conception and ten years in construction, and has far delivered a 100,000-core machine in a single 19-inch rack, which is now being expanded towards the million-core full system. Although primarily intended as a platform to support research into information processing in the brain, SpiNNaker has also proved useful for Deep Networks and similar applied Big Data applications. In this talk I will present an overview of the machine and the design principles that went into its development, and I will indicate the sort of applications for which it is proving useful.

Rudolf Kruse is Professor at the Otto-von-Guericke University of Magdeburg (Germany), where he is leading the Computational Intelligence Group. His current research interests include data science and intelligent systems. His group is successful in various industrial applications in cooperation with companies such as Volkswagen, SAP, Daimler, and British Telecom. He obtained his Ph.D. and his Habilitation in Mathematics from the Technical University of Braunschweig in 1980 and 1984 respectively. Following a stay at the Fraunhofer Gesellschaft, he joined the

Technical University of Braunschweig as a professor of computer science in 1986. Since 1996 he is a full professor at the Department of Computer Science of the Otto-Von-Guericke University Magdeburg in Germany. Rudolf Kruse has coauthored 15 monographs and 25 books as well as more than 350 refereed technical papers in various scientific areas. He is associate editor of several scientific journals. He is a Fellow of the International Fuzzy Systems Association (IFSA), Fellow of the European Coordinating Committee for Artificial Intelligence (ECCAI) and Fellow of the Institute of Electrical and Electronics Engineers (IEEE).

The title of the talk by Professor Kruse is “Modeling Self-Explanatory Big Data Applications”. Big Data and Data Science have made commercial advances driven by research. Typical questions that arise during the modeling phase are whether it should be aimed to provide explanations to the user, what should be explained and how this should be done. This talk addresses these controversies using two exemplary industrial projects: “Markov Networks for Planning” and “Medical Research Insights”.

Piotr Mirowski is a Research Scientist at Google DeepMind, the research lab that focuses on solving Artificial General Intelligence and that investigates research directions such as deep reinforcement learning or systems neuroscience. Piotr has a M.Sc. in computer science (2002) from ENSEEIHT in Toulouse, France and, prior to resuming his studies, worked as a research engineer at Schlumberger Research (2002–2005). He obtained his Ph.D. in computer science (2011) at New York University under the supervision of Prof. Yann LeCun. His doctoral work on using recurrent neural networks for learning representations of time series covered applications such as inferring gene regulation networks, predicting epileptic seizures, categorizing streams of online news and statistical language modeling for speech recognition (in collaboration with AT&T Labs Research). After his PhD, Piotr was a Member of Technical Staff at Bell Labs (2011–2013), where he focused on simultaneous localization and mapping for robotics, on indoor localization and on load forecasting for smart grids. Prior to joining Google DeepMind (2014 till now), Piotr worked as an applied scientist at Microsoft Bing (2013–2014), investigating deep learning methods for search query formulation.

The title of the talk by Dr. Mirowski is “Learning Sequences”. Many data science or control problems can be qualified as sequence learning. Examples of sequences abound in fields such as natural language processing (e.g., speech recognition, machine translation, query formulation or image caption generation) or robotics (control and navigation). Their underlying challenge resides in learning long range memory of observed data. In this talk, we will look at the inner workings of recurrent neural networks and neural memory architectures for learning sequence representation and illustrate their state-of-the-art performance.

Professors Luca Oneto and Dr. Davide Anguita, of the University of Genoa, will deliver a tutorial titled “Model Selection and Error estimation without the Agonizing Pain”. Some Big Data failures, like the infamous 2013 Google Flu Trends misprediction, reveal that large volumes of data are not enough for building effective and reliable predictive models. Even when huge datasets are available, Data Science needs Statistics in order to cope with the selection of optimal models

(Model Selection) and the estimation of their quality (Error Estimation). In particular, Statistical Learning Theory (SLT) addresses these problems by deriving non-asymptotic bounds on the generalization error of a model or, in other words, by upper bounding the true error of the learned model based just on quantities computed on the available data. However, for a long time, SLT has been considered only an abstract theoretical framework, useful for inspiring new learning approaches, but with limited applicability to practical problems. The purpose of this tutorial is to give an intelligible overview of the problems of Model Selection and Error Estimation, by focusing on the ideas behind the different SLT-based approaches and simplifying most of the technical aspects with the purpose of making them more accessible and usable in practice, with particular reference to Big Data problems. We will start by presenting the seminal works of the 80's until the most recent results, then discuss open problems and finally outline future directions of this field of research.

Professor Giacomo Boracchi, of the Politecnico di Milano, will deliver a tutorial entitled “Change Detection in Data Streams: Big Data Challenges”. Changes might indicate unforeseen evolution of the process generating the data, anomalous events, or faults, to name a few examples. As such, change-detection tests provide precious information for understanding the stream dynamics and activating suitable actions, which are two primary concerns in financial analysis, quality inspection, environmental and health-monitoring systems. Change detection plays also a central role in machine learning, being often the first step towards adaptation. This tutorial presents a formal description of the change-detection problem that fits sequential monitoring as well as classification and signal/image analysis applications. The main approaches in the literature are then presented, discussing their effectiveness in big data scenarios, where either data-throughput or data-dimension are large. In particular, change-detection tests for monitoring multivariate data streams will be presented in detail, including the popular approach of monitoring the log-likelihood, which will be demonstrated to suffer from detectability loss when data-dimension increases. The tutorial is accompanied by various examples where change-detection methods are applied to real world problems, including classification of streaming data, detection of anomalous heartbeats in ECG tracings and the localization of anomalous patterns in images for quality control.

Finally, we would like to thank the Artificial Intelligence Journal (Elsevier) for sponsoring the conference.

We hope that the 2nd INNS Big Data conference will stimulate the international Big Data community and that it will provide insights in opening new paths in Analytics by conducting further algorithmic and applied research.

September 2016

Lazaros Iliadis
Asim Roy
Marley Vellasco

Advances in Big Data

Proceedings of the 2nd INNS Conference on Big Data,

October 23-25, 2016, Thessaloniki, Greece

Angelov, P.P.; Manolopoulos, Y.; Iliadis, L.; Roy, A.;

Vellasco, M. (Eds.)

2017, XVII, 348 p. 101 illus., Softcover

ISBN: 978-3-319-47897-5