Big Data and Science Summer School
Faculty of Science and Health
The rapid and expanding provision of massive data sets, information, and technologies to interrogate these, is transforming the research landscape in science and health. This information revolution is applicable to all areas of science, from environmental data, genomic and associated information in biological systems, imaging and real time measurements, through to health data, personalised medicine, financial and empirical data, and behavioural responses (data mining of the web, blogs, twitter). All these developments are coming about through a combination of new sources of

**MONDAY 7 JULY 2014**

8.45-9.10  Registration

9.15-9.30  Welcome and introduction  
**Professor Graham Underwood, Executive Dean, Faculty of Science and Health**

**CHAIR – David Penman**

9.30-10.15  Dr Taha Yasseri, Oxford Internet Institute  
**Big Data for a new computational social science**

The use of socially generated “big data” from our daily activities has become a new technique to understand and predict our collective behaviours. Big data techniques have multifarious applications — for example predicting flu outbreaks based on the volume of tweets mentioning flu-related keywords, understanding the patterns of human mobility by analysing the records of mobile phone calls, or forecasting the financial success of a movie by studying the page view statistics of the Wikipedia articles about the movie. What all these examples have in common is the concept of quantifying and measuring activity of individuals at the collective level to understand and model human societies in a computational framework.

In this talk, after a short introduction to Big Data, few examples of the use of Big Data in computational social science will be reviewed.

10.20- 10.40 Dr Hongsheng Dai, Department of Mathematical Sciences, University of Essex  
**Challenges of big data analysis via mixture models**

Mixture models provide a statistical description of data obtained when sampling successively from randomly selected sub-populations. These models arise naturally in the areas of statistical classification and clustering. They also provide a convenient semi-parametric framework for modelling distributions of unknown shape, for example in model-based density estimation, since any distribution can be approximated by a mixture of elementary components. This talk will present an overview on the Bayesian analysis for mixture models and the challenges involved when they are applied to Big-Data analysis.
CHAIR – Berthold Lausen

11.15-12.00 Professor Claus Weihs, Technical University of Dortmund

*Big Data Analytics – Many Variables, Many Observations, and the SVM*

The talk starts with a review of definitions of Big Data Analytics. Then, it discusses the differences between Classical Data Science and Big Data Analytics. We will mainly consider two cases. In the case of many variables, we will compare dimension reduction and variables selection. In the case of many observations, we will discuss approximations based on data partitions and subsampling and (online) methods typically used for data streaming.

In the second part of the talk, a link to our current research will be given. Especially, Kernel support vector machines (SVMs) will be discussed. In contrast to linear SVMs the amount of computation required to train such a machine becomes a bottleneck when facing large datasets as in each iteration step a quadratic programming problem has to be solved. In order to resolve this shortcoming of kernel SVMs, many approximate methods for training were developed. We will compare these methods with the state-of-the-art exact solver LIBSVM on a number of large datasets. To analyze the trade-off between prediction error and runtime, we applied model-based multicriteria optimization to compute the Pareto front of these two target values.

12.05-12.25 Dr Audrey Guinchard, School of Law, University of Essex

*Big Data, privacy and data protection – An introduction to the challenges of the incoming European legal framework*

Within a year, Europe will have a new data protection framework that will leave hardly any margin of interpretation to Member States. What will be its impact on data analytics regarding its requirements of transparency and compliance with privacy rights? Where are the challenges and can we overcome them?

CHAIR – Sheina Orbell

14.00-14.45 Professor Massimo Poesio, School of Computer Science and Electronic Engineering, University of Essex

*Text Analytics for Big Data*

This presentation will cover the use of text analytics methods including information extraction/retrieval and summarization to facilitate access to and retrieval of information from big data collections.
14.50-15.35 Dr Steve Phelps, Centre for Computational Finance and Economic Agents, University of Essex

**Big Data and Computational Finance**

In this talk Steve will give an overview of agent-based modelling. Agent-based models are simulation models with an emphasis on modelling the behaviour of individuals, and are used to in many different disciplines ranging from biology, sociology and economics. Such models can be used to understand real-world complex-adaptive systems which are composed of interacting autonomous agents. A key research in this domain is how, and if, these systems maintain macroscopic homeostatic behaviour despite the fact that their constituent agents often face an incentive to disrupt the rest of the system for their own gain. The financial markets present a unique opportunity for empirically studying such systems with the recent availability of high-frequency tick-data, which records every transaction in the market, and can run to many billions of events per exchange per annum. Steve will give an overview of some of the insights that can be gained from studying these data-sets, as well as some of the technological and statistical challenges involved.

15.40-16.10 Tea

16.15-16.35 Dr Sharon Bolton, Data Curation Manager, UK Data Archive

**Dealing with Big Data: the Challenge for Archives**

The research data landscape is changing and traditional data archiving is entering a new era. This talk will focus on some of the challenges and opportunities that Big Data presents for archives: data quality assurance; ethics, confidentiality and user access to data; handling new data formats; and long-term data preservation.

16.40-17.00 Q&A/discussion

---

**TUESDAY 8 JULY 2014**

**CHAIR – Phil Mullineaux**

9.15-09.35 Professor Maria Fasli, Head of School of Computer Science and Electronic Engineering and Director of Institute for Data Analytics and Data Science (IDADS), University of Essex

**Analysing and Modelling Complex and Big Data**

This talk will be providing an overview of some of the issues and challenges arising in big and complex data from a computational point of view. We will be describing work that has been using artificial intelligence techniques such as multi-agent systems to explore and understand complex and big data in two specific strands of work. The first strand includes work on (i) modelling of financial and electronic markets such as the foreign exchange market and insights acquired through the use of real and large datasets; (ii) modelling socio-cognitive aspects such as reputation and trust in dynamic environments such as markets. The second strand of work focuses on offering personalised services to users by understanding their preferences.
and needs through their interactions with computational systems in a flexible and unobtrusive manner. Finally, some of the key challenges surrounding big and complex data and the implications on individual users as well as society overall will be touched upon.

09.40-10.25 Dr Adrian Charlton, Food and Environment Research Agency

*Finding the needle in the haystack – Targeted and non-targeted detection of food contaminants and adulterants*

10.30-11.00 Coffee

**CHAIR – Ewen Speed**

11.00-11.45 Dr Paul Taylor, Centre for Health Informatics and Multiprofessional Education (CHiME), University College London

*Issues of consent and validity in accessing and using large collections of health data*

This government is extraordinarily committed to the idea of Big Data as an engine of economic growth and specifically to the commercial exploitation of large collections of biomedical data as a way of supporting the UK’s pharmaceutical and related industries. Support for research in science and technology is seen as a useful counter to claims that their economic policies are deflationary and research that could lead to new treatments and improved health and well-being seems, at first glance, particularly benign. However, the conjunction of Big Data and Healthcare raises a number of concerns for the public, for their clinicians and for the researchers attempting to learn from the data. This talk will explore two concerns in some detail. The first is a concern about privacy. This will be discussed in the context of a stalled attempt by the government to require that GPs submit extracts from their patients’ records to a central database, care.data, which could then be used for a range of analyses by the NHS but also – in return for a fee – by commercial organizations. The second is a concern about validity and will be discussed in the context of the attempt to use mortality data as metric of the quality of care in NHS hospitals. This has seemed a successful approach, which received a great deal of positive attention in the wake of the Mid-Staffs enquiry. However, many experts remain unconvinced.

11.50-12.10 Dr Antonio Marco, School of Biological Sciences, University of Essex

*Data driven biology: do we still need hypotheses?*

Hypotheses form the backbone of modern science. In biological sciences knowledge comes from observations, yet observations alone do not produce scientific theories, and we need hypotheses and models to interpret the data. Recently, the generation of vast amounts of data has changed the way we do science. Some scholars even suggested that science is becoming ‘data-driven’, and that hypotheses are not necessary. In this talk, I will address this issue and argue that hypothesis are still needed. Indeed, many of the ‘hypothesis-free’ data analyses are actually ‘many-hypotheses’ approaches. In my opinion, this data deluge has become a data delusion, and hypotheses are still needed to produce significant progress in biological sciences.
Survival analysis on Big data: back to basics (and baselines)

When modelling complex relationships all experts call for more data. That is especially true when one tries to predict if and when an event is going to happen. It takes time to observe time -and thus more data - while in many cases the event of interest may never happen. In fact, many advanced survival models can only be applied in large data-sets with long follow up of the subjects. But there is a difference between more data and big data. Although we require large data-sets for the complex statistical models to be meaningful, most of these methods are computationally expensive. This talk will present computational challenges in large data-sets and illustrate how to fit dynamic survival models by landmarking. For even larger data-sets (=big data) we will argue that going back to the past —before the notorious Cox model was presented— can be a pragmatic solution, until we are able to buy a larger computer.

Data and other revolutions

The scientific revolution led to profound changes in our understanding of our place in the Universe. It also helped in shaping intellectual developments that led to industrial and political revolutions.

Many of the big scientific breakthroughs resulted from being able to combine data from different sources. We are now entering a new age in which we can rapidly compare data from everything with everything. This raises a number of challenges for scientists including a literature with shoddy papers and results that cannot be reproduced. I will discuss efforts to tackle these problems.

I will highlight that revolutions in data accessibility and sharing will be at the heart of many of the issues driving political change in our futures.

Big Data or Just Massive Data: Ecology in a Data Rich World

Ecology is changing. Gone are the days of the lone ecologist sitting and happily counting the number of butterflies in a field, recording their data in a notebook and easily analysing and displaying their findings on a PC. Modern ecology is big. It contains data collected using the latest technologies, from remote sensing and satellite imagery, to DNA sequencing on scales often larger than any other research area. As recently as ten years ago a terabyte of ecological data may not have been generated over the duration of a standard three-year project, yet nowadays this can be generated in an afternoon. Moreover, not only can these datasets be generated, they are increasingly being generated on an unprecedented scale. This produces two key challenges for today’s ecologists: 1) what to do with the large dataset routinely generated in their labs (massive data) and 2) what to do with all of the data
being generated across labs (truly big data). In this talk I explore the novel types of data being produced in ecology, and discuss the logistical challenges of dealing with massive data. I then explore options for dealing with the wealth of data being produced across different ecological labs and discuss how these may be combined to answer novel questions and provide new insights into ecological processes. Finally, I will examine some of the future challenges and solutions that may come into play as the size and scope of data continue to increase.

15.25-15.55 Tea

16.00-16.45 Professor Kevin Hiscock, University of East Anglia

*Advancing catchment science through high-frequency monitoring of water quality*

Traditional catchment water quality sampling by researchers and regulatory bodies has been typically limited to a daily, weekly or monthly time interval, which misses detailed information relating to storm event peaks and the ability to accurately calculate nutrient fluxes in stream runoff. With the advent of web-based sensor technology, bankside monitoring stations can record, with at least 30-min resolution, data for flow and a range of in-stream and nutrient parameters. With examples recorded in the Wensum Demonstration Test Catchment (DTC), hydrological processes and inter-seasonal and inter-annual catchment behaviour is revealed. The greater insight gained is of high value in providing evidence to policy-makers in designing mitigation measures, a number of which are being trialled in the Wensum DTC and will be introduced in this presentation.

16.50-17.20 Q&A/discussion.
50th Anniversary Celebration Meeting Classification Society

Time: 1.45pm-6.30pm
Location: Firstsite Colchester

Registration via Eventbrite at
www.eventbrite.co.uk/e/50th-anniversary-celebration-meeting-classification-society-tickets-12024251861

Speakers include:

David Hand, Imperial College London, UK
“Classifying the world: how classification methods have revolutionised understanding”

Fionn Murtagh, De Montfort University, Leicester, UK
“Data science in psychoanalysis: a short review of Matte Blanco’s bi-logic, based on metric space and ultrametric or hierarchical topology”

Maurizio Vichi, University of Rome, Italy
“Advances in multimode-clustering, clustering and dimensionality reduction”

Hans Kestler, University of Ulm, Germany
“Boolean networks for modeling biological signaling processes”

Peter Flach, University of Bristol, UK
“Recent advances in threshold selection and classifier calibration”

Contact:
Berthold Lausen, Department of Mathematical Sciences, University of Essex
E blausen@essex.ac.uk

Contact and registration:
Julie Peirson
E juliep@essex.ac.uk
T 01206 874520