THE UNIVERSITY OF WARWICK

Big Data in the Mathematical Sciences

Wednesday 13 November 2013

Sponsored by:

EPSRC

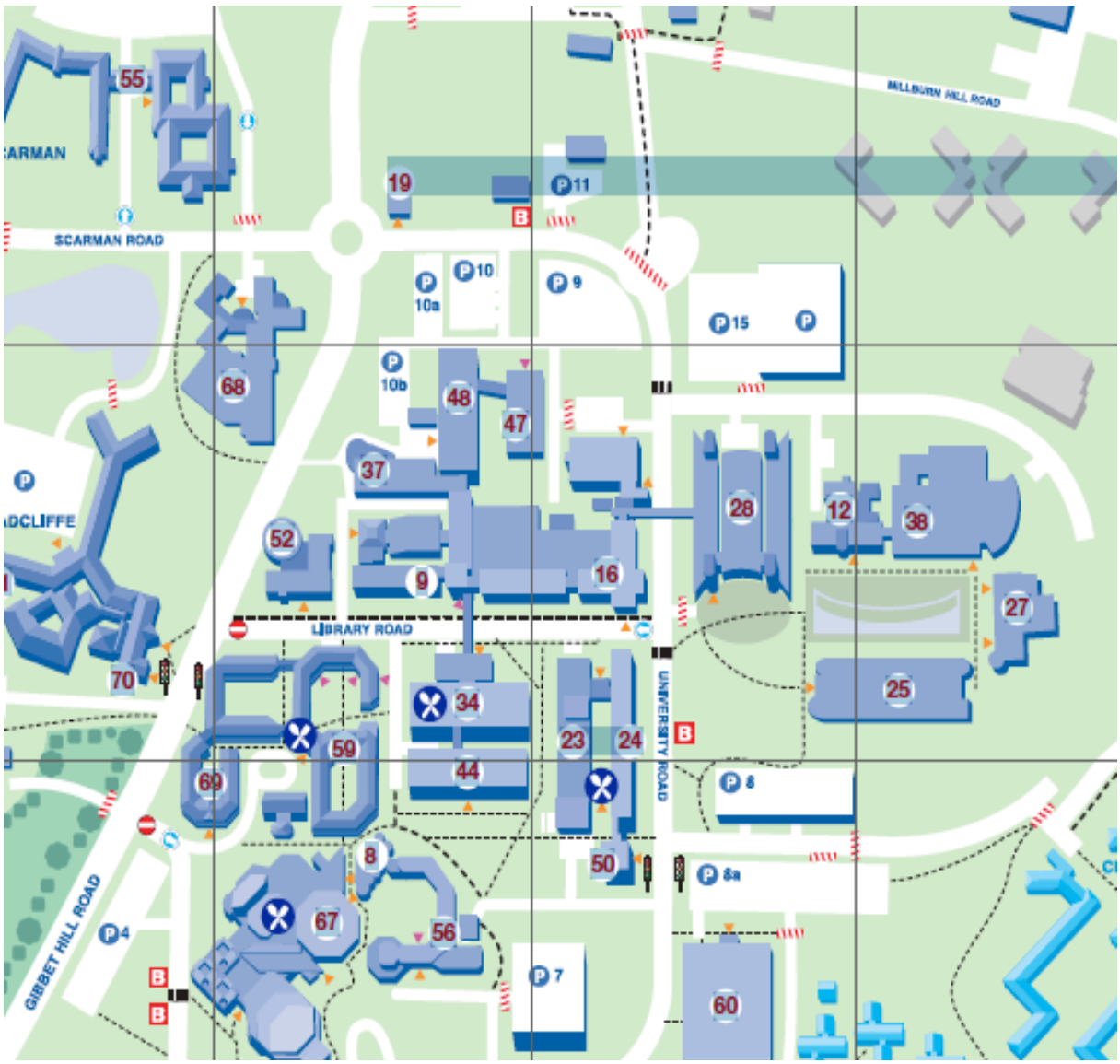
Department of COMPUTER SCIENCE

WMI

Warwick Statistics

centre for complexity science

Extract from Campus Map



Note:

Walk from Zeeman Building to Arts Centre approximately 5 minutes

Zeeman Building – Building Number 38

Warwick Arts Centre – Building Number 67

General Information

Sponsors:
The meeting is funded by the Departments of Mathematics and Statistics EPSRC Platform Grant (WAMP), with additional support from the Department of Mathematics, Departments of Statistics, Computer Science, and Complexity Centre.

Registration/Coffee
Registration will be from 10.30 a.m. on Wednesday 13 November Zeeman Building (See Map – Building 38). Coffee and refreshments will be served before first talk at 11.00 a.m.

Workshop Sessions
Morning Session – Zeeman Building (Room MS01) from 11.00 – 12.00.

Remaining Sessions - will be held at Warwick Arts Centre**,** Woods-Scawen Room (See Map - Building 67) – from 13.30 – 18.15.

Internet Access
Internet Access available to those staying on campus in their rooms, alternatively temporary user id and passwords can be obtained at registration desk.

Car Parking
Free car parking is available to all those staying on campus – on entry to car park, take blue token, this can then be validated at your accommodation reception desk. Participants who are not staying on campus will be able to park at a cost of £3.00 for a full day.

Meals
Lunch will be provided and will be served at the Warwick Arts Centre outside Woods-Scawen Room, from 12.15 – 13.30.

Special Needs:

Any participant with special needs such as mobility please contact Yvonne Carty on telephone: 02476573798 or email: y.j.carty@warwick.ac.uk.

Wednesday 13 November 2013

Coffee/Registration:	
10.30 – 11.00	Registration. Zeeman Building

Opening	
11.00 -11.05	Room MS01 – Zeeman Building

Plenary Talk: Room MS01 – Zeeman Building		
11.05 – 12.00	**Michael Jordan**	Big Data: The Computation/Statistics Interface

12.15- 13.30 - Buffet Lunch – Warwick Arts Centre

Early Afternoon Session: Woods-Scawen Room, Warwick Arts Centre		
13:30 – 14:20	**Tanya Berger-Wolf**	Analysis of Dynamic Interaction Networks
14:20 – 15:10	**Nick Duffield**	Constructing general purpose summaries of big data through optimal sampling
15:10 – 16:00	**Terry Lyons**	Reading and Learning from Time Ordered Data

16.00 – 16.30 – Coffee

Late Afternoon Session: Woods-Scawen Room, Warwick Arts Centre		
16:30 – 17:30	**Yann LeCun**	Learning Hierarchical Representations with Deep Learning
17.30 – 18.15	**Panel Discussion**	Mathematical Tools and Techniques for Big Data

Michael Jordan, University of California-Berkeley

Big Data: The Computation/Statistics Interface

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 statistical 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. We wish to blend these perspectives. Indeed, if data are a data analyst's principal resource, why should more data be burdensome in some sense? Shouldn't it be possible to exploit the increasing inferential strength of data at scale to keep computational complexity at bay? I present three research vignettes that pursue this theme, the first involving the deployment of resampling methods such as the bootstrap on parallel and distributed computing platforms, the second involving large-scale matrix completion, and the third introducing a methodology of "algorithmic weakening," whereby hierarchies of convex relaxations are used to control statistical risk as data accrue. [Joint work with Venkat Chandrasekaran, Ariel Kleiner, Lester Mackey, Purna Sarkar, and Ameet Talwalkar].

-

Tanya Berger-Wolf, University of Illinois-Chicago

Analysis of Dynamic Interaction Networks

From gene interactions and brain activity to highschool friendships and zebras grazing together, large, noisy, and highly dynamic networks of interactions are everywhere. Unfortunately, in this domain, our ability to analyze data lags substantially behind our ability to collect it. From collecting the data and inferring the networks to producing meaningful insight at scale, computational and conceptual challenges are there every step of the way.

In this talk I will show computational approaches that address some of the questions about dynamic interaction networks: whom should we sample? how often? and what are the meaningful patterns and trends? The methods leverage the topological graph structure of the networks and the size of the available data to, somewhat counter-intuitively, to produce more accurate results faster.

-

Nick Duffield, Center for Discrete Mathematics and Computer Science, Rutgers University

Constructing general purpose summaries of big data through optimal sampling

Big datasets of operational measurements have been collected and studied by internet service providers for a number of years. The amount of data presents an enormous challenge for accumulation in storage and for database management. For this reason data summarization plays an essential role, both in facilitating fast exploratory queries, and in prolonging the useful life of the data through historical snapshots. This talk shows how general purpose summaries can be constructed through a sample design that optimally mediates between the underlying data characteristics and the class of queries to be supported. This is achieved through reformulating the sampling problem in terms of minimizing a cost that combines sample size and sample-based estimation error. We illustrate this cost based approach in various settings in network measurements, discuss its computational aspects, and suggest a wider application.

Terry Lyons, University of Oxford

Reading and Learning from Time Ordered Data

Much of the important data in financial systems is streamed and multidimensional. The challenge is to make sense of, categorise, and learn from these complex data sources. Modern mathematical and statistical methods are changing our understanding of these streams and providing systematic approaches to the classification of these evolving patterns and enables the identification of functional relationships. The mathematics has evolved out of the development of the theory of rough paths and has lead to new and effective computational tools. Examples will be given relating to the classification of market streams. Interestingly, the mathematical techniques being developed here take advantage of some of the most abstract mathematics, fundamental questions about Lipchitz functions and properties of Hopf-algebras play natural roles in identifying the correct feature sets.

Yann LeCun, Center for Data Science & Courant Institute of Mathematical Sciences, New York University

Learning Hierarchical Representations with Deep Learning
Pattern recognition tasks, particularly perceptual tasks such as vision and audition, require the extraction of good internal representations of the data prior to classification. Designing feature extractors that turns raw data into suitable representations for a classifier often requires a considerable amount of engineering and domain expertise.
The purpose of the emergent field of "Deep Learning" is to devise methods that can train entire pattern recognition systems in an integrated fashion, from raw inputs to ultimate output, using a combination of labeled and unlabeled samples.
Deep learning systems are multi-stage architectures in which the perceptual world is represented hierarchically. Features in successive stages are increasingly global, abstract, and invariant to irrelevant transformations of the input.

Convolutional networks (ConvNets) are a particular type of deep architectures that are somewhat inspired by biology, and consist of multiple stages of filter banks, interspersed with non-linear operations, and spatial pooling. Deep learning models, particularly ConvNets, have become the record holder for a wide variety of benchmarks and competition, including object recognition in image, semantic image labeling (2D and 3D), acoustic modeling for speech recognition, drug design, asian handwriting recognition, pedestrian detection, road sign recognition, biological image segmentation, etc. The most recent speech recognition and image analysis systems deployed by Google, IBM, Microsoft, Baidu, NEC and others use deep learning, and many use convolutional networks.

A number of supervised methods and unsupervised methods, based on sparse auto-encoders, to train deep convolutional networks will be presented. Several applications will be shown through videos and live demos, including a category-level object recognition system that can be trained on the fly, a system that can label every pixel in an image with the category of the object it belongs to (scene parsing), and a pedestrian detector. Specialized hardware architecture that run these systems in real time will also be described.