# Machine learning meets statistics: Guiding medicine into the future

## 25-26 March 2019

**Clifford Albutt Lecture Theatre, Cambridge, UK**

## Abstracts

## Day 1 - 25 March

### Robust machine learning for causal inference in health care (Keynote)

David Sontag, MIT Computer Science and Artificial Intelligence Laboratory, and Institute for Medical Engineering and Science (USA)

Electronic health records are now pervasive, presenting an incredible opportunity to use retrospective data to learn about medicine and to improve health care. Machine learning can help answer questions such as, "What conditions does this patient have?", "When will this patient's disease progress?" and "How should we optimally treat this disease?". Properly answering these questions requires tackling head-on questions of causality, specifically how to infer causality from high-dimensional observational data. Machine learning and causal inference in health care introduces additional challenges including little labelled data, significant missing data, censoring, and the need to characterize individual-level uncertainty. I will discuss several new methodologies that my group has created to address these challenges, with a particular focus on disease progression modelling and estimation of individual treatment effect. Specifically, I discuss provable guarantees for causal inference under model misspecification (Johansson et al. ICML '16, Shalit et al. ICML '17), approaches for causal inference with unobserved confounding (Louizos et al. NeurIPS '17), how to check assumptions for off-policy reinforcement learning (Gottesman et al. Nature Medicine '19, Oberst et al. '19), assessing overlap (Johansson et al., '19), and learning nonlinear dynamical models using the deep Markov model (Krishnan et al., AAAI '17).

### Ambient intelligence in AI-assisted hospitals

Serena Yeung, Harvard University (USA)

Artificial intelligence has begun to impact healthcare in areas including electronic health records, medical images, and genomics. But one aspect of healthcare that has been largely left behind thus far is the physical environments in which healthcare delivery takes place: hospitals and assisted living facilities, among others. In this talk I will discuss my work on endowing hospitals with ambient intelligence, using computer vision-based human activity understanding in the hospital environment to assist clinicians with complex care. I will first present an implementation of an AI-Assisted Hospital where we have equipped units at two partner hospitals with visual sensors. I will then discuss my work on human activity understanding, a core problem in computer vision. I will present deep learning methods for dense and detailed recognition of activities, and efficient action detection, important requirements for ambient intelligence. I will discuss these in the context of two clinical applications, hand hygiene compliance and automated documentation of intensive care unit activities. Finally, I will present work and future directions for integrating this new source of healthcare data into the broader clinical data ecosystem, towards full realization of an AI-Assisted Hospital.

## Learning the molecular determinants of human disease trajectories

Chris Yau, University of Birmingham

The interpretation of complex high-dimensional data typically requires the use of dimensionality reduction techniques to extract explanatory low-dimensional representations. However, these representations may not be sufficient or appropriate to aid interpretation, and in many real-world problems, the physical interpretation must be made in terms of the original features themselves. Therefore characterising the relationship between latent low-dimensional representations, external covariates, and feature-level variation can be critical.

In this talk, I will describe how we can achieve this through a class of Covariate Gaussian Process Latent Variable Models (c-GPLVM) which embeds structured sparsity-inducing kernel decomposition within the GPLVM framework to allow the explicit disentanglement of feature-level variation in terms of covariate-dependent effects, contribution of latent variables, and interaction effects between the two.

I demonstrate the utility of this model for applications in disease progression modelling from cross-sectional, high-dimensional gene expression data in the presence of additional phenotypes. In each setting we show that the c-GPLVM is able to effectively extract low-dimensional structures from high-dimensional data sets whilst allowing a breakdown of feature-level variability that is not present in other commonly used dimensionality reduction approaches.

## Learning from our clinical data

Mark Baillie, David Ohlssen and Frank Bretz, Novartis (Switzerland)

Pharmaceutical drug development is the long and costly process of bringing new medicinal drugs to the market. Typically, it takes at least ten years and around $2.5 billion investment in Research & Development (R&D) for a new medicine to move from initial discovery to the marketplace. Within this journey, the use of novel technologies that enable molecular profiling, sequencing, imaging, screening, and digital clinical monitoring, have led to much faster and cheaper ways to generate huge volumes of complex and diverse biomedical data, culminating in an omnipresent challenge of big data that is new to pharmaceutical industry.

In conjunction with this data explosion, computational advances and successes from other industries have ignited interest in machine learning and artificial intelligence. The combination of the two has given rise to data science as a new discipline within pharmaceutical R&D that blends domain knowledge, computation and statistics to make the generated data useful and impactful in preclinical, clinical, manufacturing, and commercial applications. It is powered by both data analysis (e.g. computational biology/chemistry, imaging, and statistics) and data engineering (e.g. data pipelines, high-performance computing, and machine learning). Application of data science involves using mathematical foundations, statistics and coding to take very large data (unstructured and structured), clean and organize it, apply industry knowledge, analytics, and predictive models, with the ability to translate results to actionable hypotheses that address key science questions of collaborators.

In this talk, we provide some perspectives on data science in pharmaceutical R&D generally. We review preliminary experiences and learnings from our own journey so far and offer recommendations on how we can realize the promise of data sciences in future. This includes, but is not restricted to, the challenges around handling multiple and heterogeneous data sources, the need to fully embed computational and algorithmic approaches in the domain sciences, and the often overlooked need to consider the underlying causal mechanisms that gave rise to the data, rather than simply the pattern or association observed in those data.

## Human in silico clinical trials in cardiology and pharmacology

Blanca Rodriguez, University of Oxford

In silico clinical trials in medicine refer to the evaluation of a medical therapy using simulations with computer models. Already established in engineering applications (such as aeronautics), in silico trials are now starting to be more widely adopted in medicine with broad potential impact in academy, industry and regulatory bodies. The socio-economic potential in this area is thus huge. In my talk, I will describe our progress in computational modelling and simulation of the human heart towards the realisation of in silico clinical trials for cardiac pharmacology and medicine. I will describe the causes of variability in the response of human hearts to pharmacological therapy, and their importance in assessing safety and efficacy during drug development. I will then address the synergies gained from combining modelling and simulation science with machine learning and statistical approaches to unravel the causes of phenotypic variability in disease and drug response, and their implications for the advancement of human in silico trials in medicine. I will emphasize the strong collaborations underpinning this work with key partnerships in industry, regulatory agencies and experimental and clinical biomedicine. Through my talk, I will discuss the importance and challenges of inter-disciplinary, inter-sectoral collaborations in computational medicine.

### Data-driven disease progression modelling with subtype and stage inference (SuStaIn)

Daniel Alexander, UCL

My talk will introduce disease progression modelling, which aims to piece together trajectories of biomarker change in chronic disease from cross-sectional or short-term longitudinal data sets. I will discuss various approaches and applications. I will focus on the recent development of the SuStaIn (Subtype and Stage Inference) algorithm (Young Nature Communications 2018), which disentangles temporal change from phenotypic differences to identify distinct data-driven disease subtypes defined by different trajectories. I will present some recent results and discuss opportunities and challenges for future development and application.

### Cardiovascular risk prediction using big data: A statistician's perspective

Jessica Barrett, MRC Biostatistics Unit

Availability of electronic health record (EHR) data for prediction modelling offers us an opportunity to harness a wealth of data stored in patients' medical records. Focussing on the prediction of cardiovascular disease (CVD) risk, I will present a statistical approach to building a risk prediction tool using large-scale cohort and EHR data, and to the development of a more targeted approach to CVD screening and the scheduling of CVD screening.
Medical records typically contain past measurements of CVD risk factors, including blood pressure, cholesterol and smoking status. These are modelled using multivariate mixed effects models, which allow for correlations between longitudinal outcomes through correlated random effects. Patient data extracted from EHRs represents a dynamic cohort, with individuals entering and leaving the cohort over time. The lack of a natural time origin motivates a landmarking approach using age as the time-scale. In this approach, a discrete set of landmark times is specified at which risk predictions are to be made, and survival is modelled from the landmark time only for individuals still at risk. This dynamic risk prediction model allows risk predictions to be updated over time in response to new information becoming available. Finally, I will discuss the potential for additional complexities to be addressed using this approach by making adjustments to the risk prediction model.

## Day 2 - 26 March

### Multi-task time series analysis applied to drug response modelling

Chris Williams, University of Edinburgh

Time series models such as dynamical systems are frequently fitted to a cohort of data, ignoring variation between individual entities such as patients. In this paper we show how these models can be personalised to an individual level while retaining statistical power, via use of \emph{multi-task learning} (MTL). To our knowledge this is a novel development of MTL which applies to time series

both with and without control inputs. The modelling framework is demonstrated on a physiological drug response problem which results in improved predictive accuracy and uncertainty estimation over existing state-of-the-art models.

Joint work with Alex Bird and Chris Hawthorne.

### Where multi-armed bandit models meet response-adaptive randomisation for clinical trials

Sofia Villar, MRC Biostatistics Unit

Multi-armed bandit problems are an important and well-known class of models for studying the learning-earning trade-off within reinforcement learning, operations research and several other fields. Although many algorithms for these problems are theoretically well-understood and commonly used in real applications (e.g. web advertising), others remain (still) largely unused in practice. Specifically, this is the case for the extensive and rich body of literature accumulated on formulating and solving theoretical bandit models for clinical trials in which some measure of its outcome is optimized in a Bayesian setting. In such a context, the learning-earning trade-off exists between two competing goals: (1) to correctly identify the best treatment (learning) and (2) to treat as many patients as effectively as possible during the trial and after (earning).

In this talk, I review some important results from this literature, present their advantages and limitations to their use for designing clinical trials in practice and discuss how bandit algorithms can be viewed (and modified) to define response-adaptive randomisation procedures. I will illustrate how different bandit algorithms perform in terms of different operating characteristics in the context of real clinical trials. I will also discuss how this line of work has resulted from an attempt of bringing theory closer to practice while providing an answer to the current challenges of therapy development for rare diseases and personalised medicine.

### Deep reinforcement learning: Challenges and successes (Keynote)

Oriol Vinyals, Google DeepMind

Deep Reinforcement Learning has emerged as a sub-field in machine learning which extends the capabilities of Deep Learning systems beyond supervised and unsupervised learning. In the last few years, we have witnessed advances on domains in which complicated decisions must be carried by an "agent" interacting with an "environment". In this talk, I will summarise the state of deep RL, highlighting successes from ATARI, to Go and StarCraft, as well as depicting some of the challenges ahead.

### Low-priced lunch in conditional independence testing

Rajen Shah, University of Cambridge

Conditional independence testing is central to a number of statistical problems such as variable selection, graphical modelling and causal inference, to name a few. We show however that any test with correct size does not have power against any alternative: in this sense conditional independence testing is a statistically impossible task.

Given the non-existence of uniformly valid conditional independence tests, we argue that tests must be designed so their suitability for a particular problem setting may be judged easily. To address this need, we propose to nonlinearly regress X on Z, and Y on Z and then compute a test statistic based on the sample covariance between the residuals, which we call the generalised covariance measure (GCM). We prove that the validity of this form of test relies almost entirely on the weak requirement that the regression procedures are able to estimate the conditional means X given Z, and Y given Z, at a slow rate; that is they are able to predict well. While our general procedure can be tailored to the setting at hand by combining it with any machine learning method for regression, we develop the

theoretical guarantees for kernel ridge regression. A simulation study shows that the test based on GCM is competitive with state of the art conditional independence tests.

### Unsupervised learning in high dimensions via adaptive projections

Sach Mukherjee DZNE, (Bonn)

How should we deal with latent group structure in very high dimensions, when the groups might differ not only in their location but in group-specific covariance/graphical model structures? This question is relevant in applications, for example in detecting disease subtypes, and in scientific interpretation, due to the implications of Simpson's paradox. Despite much recent progress on mixtures of high-dimensional graphical models, the very high-dimensional case remains statistically and computationally difficult. I will discuss some recent work that brings together projection ideas that have been extensively investigated in machine learning with model-based clustering as long studied in statistics. I will present methodology that revisits the problem from an assignment risk point of view, leading to an adaptive scheme that is both demonstrably effective and highly scalable.