ADVANCES IN DATA SCIENCE 2019

Accepted Posters

1. **Sumon Ahmed,** Alexis Boukouvalas, Magnus Rattray

Uncovering gene-specific branching dynamics with a multiple output branched Gaussian Process (mBGP)

Identifying branching dynamics from high-throughput single-cell data can help uncover gene expression changes leading to cellular differentiation and fate determination. Boukouvalas et al. (Genome Biology 2018) recently developed a branched Gaussian Process (BGP) method that provides a posterior estimate of gene-specific branching time with associated credible regions. Their method generalises the Overlapping Mixture of Gaussian Process (OMGP) model (Lazaro-Gredilla et al, Pattern Recognition 2012) by using a branching kernel function to account for dependence between the Gaussian Process functions in the mixture model. However, inference in this model is performed independently per gene, resulting in two significant drawbacks. Firstly, the model suffers potentially inconsistent cell assignment, i.e. the same cell may be assigned to different branches for different genes. This risk can be mitigated by using strong global priors for the cell assignments (identified by a global method such as Monocle) but when analysing pioneer genes, no cell branch prior assignments are available and this leads to potentially inconsistent cell assignments. Secondly, the computational requirements are very high when analysing many genes, making the model impractical for large numbers of cells and genes. We propose a multiple output branching Gaussian Process (mBGP) model that performs inference jointly across all genes of interest. This approach solves the inconsistency issue as well as being more computationally efficient because cell-assignments are shared across genes. Our approach involves two main ideas: (1) We develop a joint model with a different branching time parameter for each output dimension (gene). This allows for consistent estimation of the cell allocation across all genes as the cell allocation is shared for all genes, and is also more computationally efficient as there are less parameters to learn. (2) We develop a gradient-based approach to learn branching times. Using gradients removes the need for a grid search, which is impractical in the multiple-gene case since there is a combinatorial explosion of the number of branching time combinations. To verify our approach, we have applied it on both synthetic and real single-cell RNA-seq gene expression data. We show that the model can jointly estimate all branching times with significantly less computational time compared to the original BGP model, whilst also ensuring cell assignment consistency.

2. **Ghada Alfattni,** Goran Nenadic, Niels Peek

Integrating Drug-exposure Information from Structured and Unstructured Patients' records in EHRs

Electronic health records (EHRs) contain a wealth of routinely collected data that is key for understanding patient treatments and disease patterns. Apart from structured data about diagnoses and biomarkers, these records often include unstructured free-text data such as clinical

		notes, radiology reports and free-text discharge summaries which include detailed information about hospital course, treatments provided and medication prescriptions. These two data types often provide complementary information. However, disjoint analytics toolsets exist for structured and unstructured data, making it difficult to analyse datasets that comprise both types of data. In this study we explore extracting, integrating and representing drug-exposure data and their temporal relations in various sources in EHRs, using the MIMIC III dataset. The data includes drug exposure duration, dosage, frequency of application, mode etc. In addition to the identification of all attributes, the main challenges include identification of temporal relations in unstructured notes and event coreference resolution across multiple sources of EHRs (i.e., across structured and unstructured data). We will also demonstrate how using structured data along with information extracted from free-text data is useful for informing clinical decisions, data modelling and pharmacovigilance research.
3.	**Fatima Almaghrabi,** Prof Dong-Ling Xu, Prof Jian-Bo Yang	***A Comparative Study on Feature Selection for Trauma Outcomes Prediction Models*** Various demographic and medical factors have been linked with mortality after suffering from traumatic injuries such as age and pre-injury comorbidities. A considerable amount of literature has been published on the building of trauma prediction models. However, few analyse the features selection criteria. Patient records comprise a large amount of data and numerous variables, and some are more important than others. Highlighting the most influential variables and their correlations would assist in the better use of them. The intention of this study is to clarify several aspects of demographic and medical factors that could affect the outcome of trauma in order to exhibit the interaction between these factors and to represent their relationships. In addition, the aim is to use ranking and weights of features to select the features that increase prediction accuracy and lead to better results.
4.	**Mohamed Bader El-Den**	***A Biased Ensemble Approach For Dealing With The Class Imbalance Problem*** This talk presents a new method for dealing with the class imbalance problem named Biased Random Forest (BRAF). The algorithm is motivated by the idea of moving the oversampling from the data level to the algorithm level. In other words, instead of increasing the minority instances in the dataset, the algorithm in this paper aims to "oversample the classification ensemble" by increasing the number of classifiers that represent the minority class in the ensemble. This oversampling of the classification ensemble aims to generate an ensemble biased towards the minority class to compensate for its low presence in the dataset. BRAF consists of three main stages. Firstly, the nearest neighbour algorithm is employed by the BRAF algorithm to identify the difficult/critical areas in the dataset, which are the minority instances and their k nearest majority neighbours. Secondly, a standard random forest is generated from all the records in the dataset. Thirdly, the standard random forest is then fed with more random-trees generated based on the difficult areas, hopefully resulting in a more diverse

ensemble/forest and at the same time, biased towards the minority class. The bias in the forest aims to overcome the low presence of instances belonging to the minority class(es). The performance of BRAF has been evaluated on both real-world and artificial binary imbalanced datasets and compared against other state-of-the-art methods. The results showed that BRAH has improved the performance of the base random forest classifier in terms of performance and diversity. Also, BRAH has outperformed other methods such as SMOTE (Synthetic Minority Over-sampling Technique) on several datasets.

5. **Mohamed Bader El-Den,** Aya Awad, James McNicholas, Jim Briggs

Towards Early hospital mortality prediction in intensive care units – A Data Mining Approach

Mortality prediction of hospitalized patients is an important problem. Over the past few decades, several severity scoring systems and machine learning mortality prediction models have been developed for predicting hospital mortality. By contrast, early mortality prediction for intensive care unit patients remains an open challenge. This study highlights the main data challenges in early mortality prediction in ICU patients and introduces a new machine learning based framework for Early Mortality Prediction for Intensive Care Unit patients (EMPICU). The proposed method is evaluated on the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database. The results show that although there are many values missing in the first few hours of ICU admission, there is enough signal to effectively predict mortality during the first 6 h of admission. The proposed framework, in particular, the one that uses the ensemble learning approach – EMPICU Random Forest (EMPICU-RF) offers a base to construct an effective and novel mortality prediction model in the early hours of an ICU patient admission, with an improved performance profile.

6. **Michael Barrowman,** Glen Martin, Mark Lambie

Development of a Multi-State Clinical Prediction Model in Chronic Kidney Disease using the Salford Kidney Study Dataset

Introduction: To develop and compare multi-state clinical prediction models (MSCPMs) for death and/or renal replacement therapy (RRT) for patients suffering from chronic kidney disease (CKD).The methodology surrounding the specific development of a MSCPM is sparse and so methods from other modelling techniques have had to be hijacked and altered to be applicable to this field. Design: Two-, Three- and Five-State clinical prediction models were developed in a cohort of CKD patients from the Salford Kidney Study (SKS) in the Renal Department of Salford Royal NHS Foundation Trust (SRFT). 2,971 patients aged 18+ with Stage 3+ CKD were recruited between October 2002 and December 2016. They were monitored for Death, Death with or without RRT and Death with or without RRT stratified by modality. Results: 1,413 patients died, 672 began RRT (HD:339, PD:228 and Tx:105). Discussion: The three models can provide prediction for patient outcomes at any time point in the future up to 12 years. By developing and comparing the predictive performance of the models to eachother, Can we establish which would be the best to implement in clinical practice?

7. **Daniel Bloor**

Towards a Real-time Monitoring Platform for Self Management of

Chronic Obstructive Pulmonary Disease

The British Lung Foundation claims that, in 2012, Chronic Obstructive Pulmonary Disease (COPD) was affecting an estimated 1.2 million people in the United Kingdom alone, resulting in approximately 30,000 deaths per year. As research and technology progresses, new opportunities to monitor the condition of chronic lung disease sufferers, both for themselves and for the healthcare provider supporting them. Our project aims to develop a monitoring platform for COPD sufferers to self-manage their symptoms and healthcare providers to monitor the patient away from the clinic, to forecast the development of the symptoms, and to apply predictive analytics for evaluation the user's risk of exacerbation. Current developments involve research into sensor technologies and their application for COPD monitoring, including ECG for respiratory rate identification, and air quality monitors for understanding environmental contexts. The data is used for assisting in developing a processing and visualisation workflow which will then be implemented as part of a mobile application with additional interface features including providing biofeedback and requesting user context to aid in providing personalised support. The project is working towards a comprehensive platform which can collect environmental context, daily user context, physiological readings and clinical context to return exacerbation risk through personalised, explainable predictive analytics and symptom forecasting. Personalised suggestions to the user will encourage lifestyle improvements and prompt early intervention to prevent emergency hospital visits and symptom development.

8. **Michael P.J. Camilleri,** Christopher K.I. Williams

A Model for Learning Across Related Label Spaces

We consider the problem of carrying out learning in situations when classes or features are not consistently annotated with the same schema. For example, in supervised document classification, expert annotators may be asked to fully specify the topic to which a document belongs if it falls within the domain of their expertise, while only indicating the general field otherwise. Our contribution is to show that even if different annotators use disparate, albeit related, label sets, then we can still draw inferences for the underlying "complete" label set. We propose the Inter-Schema AdapteR (ISAR) to translate the fully-specified label set to the one used by each annotator. Our approach enables learning under this condition of heterogeneous schemas, without the need to re-annotate the data. Furthermore, by being able to use data from multiple schemas, ISAR is able to boost classification performance significantly over just using the limited data in one schema. In a document labelling scenario we observed an increase by a factor of 34% in F1-score, while applied to a synthethic data-set of crowd-sourced multi-annotations, ISAR showed a 14% relative increase in accuracy. Moreover, the marginal observed log-likelihood on a real crowdsourced behavioural characterisation data-set increased from -3.49 to -2.39.

9. **Adam Farooq,** Yordan Raykov, Luc Evers, Max Little

Adaptive probabilistic principal component analysis
Using the linear Gaussian latent variable model as a starting point we relax some of the constraints it imposes by deriving a nonparametric latent feature Gaussian variable model. This model introduces additional discrete latent variables to the original structure. The Bayesian nonparametric nature of this new model allows it to adapt complexity as more data is observed and project each data point onto a varying number of subspaces. The linear relationship between the continuous latent and observed variables make the proposed model straightforward to interpret, resembling a locally adaptive probabilistic PCA (A-PPCA). We propose two alternative Gibbs sampling procedures for inference in the new model and demonstrate its applicability on sensor data for passive health monitoring.

10. **Robert Firth**

Automatic Annotation of Protein Residue Mentions in Published Papers
There are 144042 Biological macromolecular structures in the Protein Data Bank, a key resource for many life-science researchers in both Industry and Academia. For many proteins, there is more than one structure in the PDB, and each of these structures may contain hundreds, or even thousands of amino acid residues. For researchers who may be interested in specific sites of interaction, mutation or substitutions that may only be a few or even a single amino acid long, finding relevant publications can prove overwhelming, with around 10,000 relevant papers published each year. We present a new text mining tool, 'pyresid' for mining full text articles and abstracts for specific mentions of amino acid residues and associated annotations. Pyresid provides a bridge between the various data sources, a sophisticated information extraction algorithm, and a link to cutting edge Natural Language Processing (NLP) pipelines, opening up opportunities for further intelligent work based on a seed-set of high-precision/low-recall pattern-matched annotations. Pattern matching has been used in the process of text mining similar corpora, two developments have made it possible to improve on this previous work. Firstly, fast open source Natural Language Processing (NLP) pipelines are now available, and can provide processing speeds orders of magnitude above those previously used. Secondly, the EuropePMC and the PDB in Europe (PDBe) provide APIs to their repositories of manuscripts which allow integration of the NLP tool developed with these data sources. Our pipeline provides as an output amino-acid level annotations that are soon to be integrated into an online platform but can also be used standalone as an powerful open-source toolkit.

11. **Juan Jose Giraldo Gutierrez,** Pablo Moreno Munoz, Mauricio Alvarez

Natural Gradient Optimisation for Improving Inference of Heterogeneous Multi-output Gaussian Process Models
In the context of large-scale Gaussian processes (GP) and their sparse approximations is growing an interest in multi-output models and their direct application to real-world scenarios (e.g. health sciences or meteorology). It has been experimentally proven that by exploiting the underlying correlations between multiple outputs, it is possible to obtain better predictions rather than considering independent tasks. However, in most cases, the literature has focused on multiple regression problems for continuous variables, generally assuming Gaussian likeli- hoods.

Recently, regarding the statistical types to be modelled, Moreno-Muñoz et al. (2018) presented their heterogeneous multi- output Gaussian process (HetMOGP) model. The main issue this this type of model comes from the strong conditioning between variational distributions and hyper-parameters of the HetMOGP prior making it extremely sensitive to any small change on any of the parameters. with the aim to overcome the issues present in a sort of model like the HetMOGP and improve inference over the set of hyperparameters, we borrow ideas from variational optimisation defining an exploratory Gaussian distribution. Moreover we use the mirror descent method in the mean-parameter space so as to keep exploiting the NG method not only over each variational posterior, but also over that exploratory distribution over the remaining variables.

12. **Simon Goodchild,** Louise Butcher

Predicting Care Pathways following hospital discharge

Predicting Care Pathways following hospital discharge (proposed poster abstract) An important problem for NHS Central Commissioning Groups is to know what happens to emergency patients when they are discharged from hospital. Patients often need post-admission care, and cannot safely be released until this is ready. A predictive model for pathways can allow CCGs to use their services more efficiently, and stop patients from suffering delayed discharges. The STFC Hartree Centre's Data Science Team worked to build predictive models for these care pathways. The CCG provided suitably anonymised data sets, covering GP and hospital care, social care, mental health and community services for patients in Liverpool. Each of these disparate data sets needed to be cleaned, checked for anomalies and mistakes and then joined together to create a single profile for each patient. This took significant effort, but created for the first time a fused data set of detailed medical histories which could be used to train predictive models. Visualising these histories and the pathway statistics provided the CCG with useful information about the usage of different services and the way patients move through the system. Using multivariate logistic regression and random forests, we were then able to train models which have about 50% accuracy for directly predicting care pathway in a balanced sample taken from the patient population. With the number of different pathways available, this is a substantial improvement over the CCG's previous knowledge.

13. **Jason Gofford**

Understanding commercial radio listener behaviour

Peak have been working with Global, Europe's largest commercial radio organisation, to enable data-driven decisions. Decision-making in broadcast radio has traditionally relied on expensive and time consuming in-person user questionnaires and sampled listenership testing. Results from these sampled tests can often have turn-around times in the region of months which makes it difficult to make reactive decisions. The proliferation of digital listening in recent years — enabled in part by, for example, the boom in smartphone and in-home smart-assistant usage — has given broadcasters like Global access to enormous amount of rich data about their listeners and their listening behaviour. This listener data, combined with metadata about about broadcast content, finally

enables the radio broadcaster to understand how their broadcast content affects listenership in near-real-time. Peak are currently working with a 2-year-long continuous transcript for content broadcast across the Global radio network. Focussing on the transcript for Capital London, containing every distinct word broadcast across songs, adverts and presenter links on the station since February 2017, Peak have applied NLP techniques to help Global not only identify their presenter-based content but also understand how that content resonates with their online listenership. In this talk we will discuss how we isolate and remove pre-recorded ad- and song-based content from the corpus of continuous text, and subsequently show how we use Latent Dirichlet Allocation (LDA) and Labelled Latent Dirichlet Allocation (L-LDA) to establish a performance baseline for the various categories of presenter link content. We discuss how a data-driven approach to broadcast content can be used by commercial radio stations to optimise their content to maximise listenership retention.

14. **Carlos Gonzalez Zelaya**

Towards Explaining the Effects of Data Preprocessing on Machine Learning

Ensuring the explainability of machine learning models is an active research topic, naturally associated with notions of algorithmic transparency and fairness. While most approaches focus on the problem of making the model itself explainable, we note that many of the decisions that affect the model's predictive behaviour are made during data preprocessing, and are encoded as specific data transformation steps as part of pre-learning pipelines. Our research explores metrics to quantify the effect of some of these steps. In this initial work we define a simple metric, which we call volatility, to measure the effect of including/excluding a specific step on predictions made by the resulting model. Using training set rebalancing as a concrete example, we report on early experiments on measuring volatility in the Students' Academic Performance dataset, with the ultimate goal of identifying predictors for volatility that are independent of the dataset and of the specific preprocessing step.

15. **Lameiece Hassan,** Mahmoud Elhawati, Mary Tully, James Cunningham, Miguel Belmonte, Goran Nenadic

#Datasaveslives: mixed methods analysis of a social media campaign to promote the benefits of using health data for research

The microblogging platform Twitter (www.twitter.com) has become popular among scientific researchers1,2. Twitter allows users to post short (previously 140 character, more recently extended to 280) messages, known as 'tweets', which may include URL links, multimedia content (e.g. images or videos) and/or references to other users (signified using the '@' symbol, plus a username). Hashtags may also be used by assigning the '#' character to a term of their choice; this is a useful way of indexing and finding tweets on a similar topic. The wider literature indicates scientists use Twitter to disseminate formal scientific work, engage with other scientists and communicate with broader audiences, including the public3,4. Whilst there is increasing use in why and how individual academics use social media, there is less research concerning how it is used as part of higher level strategic communications by departments and organisations. One example of

such communications in the field of health informatics is the public engagement campaign #datasaveslives. Developed by the Farr Institute in 2014, the campaign uses a simple hashtag to promote the positive use of data in health research on social media. Supporters of health data research are encouraged to use the hashtag #datasaveslives on social media sites (primarily Twitter) to index examples of research and wider online content that demonstrate how health data from patient records and other sources can be used to create public health benefits.

16. **Thomas House**

How are you Feeling?

One of the key possibilities opened up by modern data science is the ability to track human feelings and behaviour. I will present a set of methodological approaches to this problem based on the mathematics of probability that give insights into:

- Friendship protecting you from depression
- Life satisfaction over time (it gets better!)
- Chronic pain and mood
- How memes spread online, and why Mark Twain was right

17. **Rebecca Howard,** Danielle Belgrave, Panagiotis Papastamoulis, Angela Simpson, Adnan Custovic, Magnus Rattray

Development of allergic response data through childhood

The temporal patterns of allergic sensitisation can now be inspected at a greater resolution as a result of Component Resolved Diagnostics (CRD) (Treudler and Simon, 2013). We are using this data to group similar allergic sensitisation patterns with the aim is to better understand co-sensitisation patterns, relate them to allergy-related disease risk, and inform the development of a more personalised approach to disease diagnosis, management and treatment. Here, we scale up previous analyses of immune response data (component-specific IgE (sIgE)) to small subsets of the available allergen data by considering the patterns of response to allergens from all available sources, and across six time points from infancy to adolescence. We measured sIgE immune responses in participants from a well-characterised population-based birth cohort at six follow-ups between the ages of 1 to 16 years (ages 1, 3, 5, 8, 11 and 16). We used a Bernoulli mixture model with a Bayesian MCMC algorithm to learn clusters of sIgE components from binarised sensitisation data, i.e. each cluster contains allergen-related components with a similar sensitisation profile across the children. Model parameters and optimal number of clusters were inferred at each age. The flow of allergens between clusters across time is shown in Figure 1, showing clear and consistent patterns in the data, and allergens cluster into increasingly specialised groups according to associated child responses. Though each age was clustered independently of the others, the clusters were biologically meaningful, had exceptionally high mean assignment probabilities, and -- as Figure 1 shows -- displayed a high degree of consistency and stability across time points. The cluster-based sensitisation profiles of participants across these ages were then related to asthma and hay fever variables at age 16. When subject responses are stratified appropriately (taking into account the heterogeneous nature of both the subjects and the diseases themselves), the allergic response at age 5 can be strongly associated with the development of asthma and hay fever at age 16. We identified combinations of cluster, time point and degree of cluster sensitisation that were clearly linked to an

increased risk of asthma and hay fever development (an example of which is shown in Figure 2), as well as putative "lead" components (e.g. Fel d 1, from cat). Further application of this Bayesian clustering approach to similar data, and the continued exploration of the resulting clusters ought to facilitate the development of better diagnostic and prognostic biomarkers for allergic diseases.

18. **Glorianna Jagfeld,** Steven Jones, Paul Rayson, Fiona Lobban

How do people describe personal recovery experiences in bipolar disorder in structured and informal settings?

Bipolar disorder is a severe mental health condition characterised by changing episodes of intense depressed and elevated mood. While these symptoms are clinically regarded as chronic, modern mental health research posits that personal recovery is possible: living a satisfying and contributing life alongside symptoms of severe mental illnesses. To date, personal recovery in bipolar disorder has been investigated qualitatively with structured interviews and quantitatively with standardised questionnaires of mainly English-speaking westerners. This PhD project aims to broaden this scientific evidence base by incorporating evidence from unstructured settings and a sample of individuals from more diverse cultural and ethnic backgrounds. Therefore, we will collect and analyse textual social media data that relate to personal recovery in bipolar disorder from Twitter, the discussion forum Reddit, and blogs. Target users on Twitter and Reddit can be identified by matching self-reported diagnosis statements such as 'I was diagnosed with bipolar disorder'. Corpus and computational linguistic methods allow us to efficiently analyse these large-scale data. We will start with exploratory quantitative research using comparative corpus analysis tools to uncover important linguistic features, e.g., keywords and key concepts that occur with unexpected frequency in our collected datasets relative to reference corpora. This will be complemented by computational linguistic methods such as topic modelling and sentiment and emotion analysis as well as qualitative, manual investigations of fewer examples. We will compare and relate our insights to those of previous qualitative research conducted with traditional interviews. Since mental health constitutes very sensitive information, ethical considerations are importantimportant, and the proposal is reviewed by the departmental research ethics committee. Informed consent will be sought whenever possible, which is infeasible on Twitter and Reddit, but applicable in the case of manually selected blogs. To protect the anonymity of the users, we will paraphrase all social media quotes in our publications because usernames could be retrieved via web searches otherwise. Throughout the project, we will consult a panel of individuals with lived experience of bipolar disorder to guide our research and discuss ethical considerations. The results of this project will allow us to draw a more completer picture of the facets of personal recovery in bipolar disorder and the factors that facilitate or hinder it. This has direct implications for the design of mental health services, which we believe should be informed by voices of individuals as diverse as those they are supposed to serve***.***

19. **Stefanos Kollias,** James Wingate, Ilianna Kollia, Luc Bidaut

Healthcare Prediction using Latent Information extracted from Deep Neural Networks

Deep learning and deep neural networks have managed to produce remarkable improvement of achieved performances in analysis of multimedia information during the last few years, sometimes surpassing human perception capabilities. It is, therefore, evident that their use in the human health care and well being fields can produce significant advances in disease early diagnosis, risk prevention, assisting, all, but especially older persons to achieve a high quality of life. However deep neural networks lack the required transparency in their decision making and the on-line adaptation capability. This makes their use difficult in fields such as healthcare, where trust and personalised treatment are of high importance. In this paper we present a new methodology, for disease prediction, that possesses transparency, person and domain adaptation capabilities. We start with state-of-the-art training of deep convolutional and convolutional – recurrent neural networks to classify their input data in two classes, i.e., healthy subjects and patients. We then extract latent variable information from various levels of the trained networks and appropriately cluster them, deriving more detailed representations of the disease under investigation. By (medically) annotating the clustered representations, we generate a set of cluster centroids, together with the respective input data and annotations, which can represent different stages of the disease. This set is then used in three scenarios: a) to transparently predict new subjects' status; b) to retrain the networks with new subjects' data; c) to transfer the achieved learning to other similar environments, using a domain adaptation technique. The health care application we use to illustrate the capabilities of the proposed approach is prediction of Parkinson's, that is a neurodegenerative disease that highly affects, especially older, persons nowadays. Our developments are tested on medical imaging data, DaT Scans and MRI data, from a recently developed large Parkinson's database. Excellent results are reported related to all the above mentioned scenarios.

20. **Tristan Millington,** Mahesan Niranjan

Gene Regulatory Network Inference via Cardinality Constrained Optimisation

Understanding the interactions between genes and how these interactions change during disease is an approach that can help us improve our knowledge of said diseases and develop new treatments. We explore an efficient newly developed method that constraints the number of non-zero values in a regression problem to infer partial correlation networks. Applying this method to the TCGA Breast Cancer dataset we study genes whose regulation is altered by the cancer**.**

21. **Hari Mohan Pandey**

A Modified Whale Optimization Algorithm with Multi-Objective Criteria for Optimal Robot Path Planning

Exploration and exploitation are the two important property of every search and optimization algorithm. Exploration aims to visit entirely new region of a search space whilst, on the other hand exploitation focuses on those regions of a search space recently visited. To be successful, optimization algorithms need to setup a proper mechanism to achieve

good exploration and exploitation. Whale Optimiza-tion Algorithm (WOA) is a nature-inspired metaheuristic algorithm that mimics the hunting behavior of humpback whales. WOA achieves both exploitation and exploration respectively through bubble-net attacking method and search for prey. In this paper, we present a Modified Whale Optimization Algorithm (MWOA). It includes two additional parameters: whale memory and a new random search agent. Whale memory is used to enhance exploration whilst, on the other hand a random search agent is used to improve the exploitation capability of the WOA. The performance of the proposed MWOA is studied extensively. To do so, we implemented MWOA for robot path finding problem. Numerical results and comparisons are then shown to highlight the effectiveness and superiority of the proposed MWOA.

22. **Georgiana Neculae,** Gavin Brown

Ensembles of Spiking Neural Networks

The goal of ensemble learning is to combine multiple predictions of the same phenomenon with the aim of better representing the phenomenon. This has been shown to generally happen when the individuals are diverse and effectively combined. There is increasing evidence that ensemble systems are generally used in the brain. As mathematical models capable of replicating many computational dynamics of biological neurons. Spiking neurons are a natural substitute of biological neurons to be used in simulations. Thus, understanding how the information from several spiking neurons should be combined is an important step towards understanding how biological neurons collaborate while at the same time improving the simulation performance of spiking networks. We propose a theoretical study of the theoretical principles of combining the predictions of a group of spiking neural networks.

23. **Nicole Nisbett**

Harnessing Citizen Input: A new tool for Parliaments

There has been a concerted effort within the UK Parliament to increase their digital outreach and explore ways in which the internet can be used to aid democracy (Digital Democracy Commission 2015; Liaison Committee, 2015). While the use of social media and online petitions has increased (for example the recent RevokeArticle50 petition with over 5 million signatures), digital engagement in the form of online consultations with the public have created issues around high volumes of textual data. This poster will introduce a solution to these problems relating to parliamentary digital engagement activities. This involves a collaborative, purpose-built web-application which provides visualisations of the data from online discussions, and combines elements of natural language processing to provide a condensed overview of a debate. Word and bigram frequencies (Aggarwal and Zhai, 2012), topic modelling (Blei, Ng, and Jordan, 2003; Hong and Davidson 2010), sentiment analysis (Neilsen, 2011; Mohammed and Turney, 2013), and readability scores (Flesch, 1948) are used to provide an outline of particular themes and citizen suggestions which are most prominent in the online discussion, how the participants feel about those themes, and some insights into the demographics and socio-economic background of the participants. We have tested other topic model approaches to the

popular Latent-Dirichlet Allocation (LDA) including using bigrams instead of unigrams to improve to the interpretability of the topics. Where possible, these methods of natural language processing are combined with social network analysis to uncover how the participants interact with each other in a particular discussion, i.e. whether there are opposing interest groups. We have found that Facebook users cluster depending on the topic of post they have engaged with, and that certain topics attract more focussed and enthusiastic users. These insights allow professionals in Parliament to take advantage of natural language processing techniques and contribute to their work, in a way that was not possible before. The visualisations have been chosen specifically for being easy to understand by people from a non-technical background, to make the application of data science more inclusive and valuable. This frees up time and staff resources to concentrate on other tasks and encourages officials to conduct more digital engagement activities, ultimately allowing the public to participate in political discussions with Parliament knowing their comments will be taken into account.

24. Matiss Ozols, Alexander Eckersley, Sarah Hibbert, Jerico Revote, Jiangning Song, Christopher Griffiths, Rachel Watson, Mike Bell Michael Sherratt

From unstructured scientific peer reviewed literature to novel biomarkers of ageing.

Average human lifespan has increased substantially over the last decades; however, ageing is still an unavoidable, natural process which is associated with pathological tissue remodelling as a result of passage of time and exposure to environmental factors. Human skin is a good model organ in which to study ageing. Not only is there a cosmetic and dermatological interest in rejuvenating skin and studying cutaneous pathologies but the constituent tissues are accessible via minimally invasive biopsies and subject to environmental stressors. This interaction can lead to protein degradation and the release of a protein fragments (matrikines) that have been showed to influence cell phenotype and, in the form of topical treatments, to induce rejuvenation of tissue structure and function. In order to understand age-related changes in an organ such as skin it is necessary first to establish the baseline young, healthy proteome. However, we have shown that there is little consensus between existing experimentally determined skin proteomes. Therefore, in order to define the healthy skin proteome (Manchester Skin Proteome: MSP) we developed a novel approach which combined text mining, web-scaping and systematic literature review approaches to screen peer-reviewed publications in the 'Web of Science' and 'Pubmed' databases. The resultant, consensus skin proteome is hosted on http://www.manchesterproteome.manchester.ac.uk/. We have subsequently used this proteome to identify potential protein biomarkers of ageing. Proteins were stratified according to known degradative mechanisms: UV, ROS, glycation, protease-mediated proteolysis and DNA damage. This analysis predicts that, in general, collagens will be degraded by extracellular proteases whilst elastic fibre associated proteins will be susceptible to UVR and ROS. In order to validate these predictions mass-spectrometry proteomics methods have been employed to distinguish proteomic differences between photoaged and intrinsically aged skin. In order to further refine this analysis we have developed a new bioinformatic approached to predict regional

susceptibility within protein structure and to validate these predictions by mapping peptide fingerprints within mass spectrometry data. To stimulate software sustainability a novel structural proteomics analysis web-application is under development to allow users to analyse their sequences of interest to with regards to stressors and to map changes in peptide fingerprints within mass spectrometry datasets.

25. **Benjamin Puitong Lam,** Paolo Missier

P4@NU: Exploring the role of digital and genetic biomarkers to learn personalized predictive models of metabolic diseases

Medicine is evolving, from the inefficient detect-and-cure approach and towards a Preventive, Predictive, Personalised and Participative (P4) vision that focuses on extending individual wellness state, particularly those of ageing individuals [1], and is underpinned by "Big Health Data" such as from genomics and personal wearables [2]. P4 medicine has the potential to not only vastly improve people's quality of life, but also to significantly reduce healthcare costs and improve its efficiency. The P4@NU project (P4 at Newcastle University) aims to develop predictive models of transitions from wellness to disease states using digital and genetic biomarkers. The main challenge is to extract viable biomarkers from raw data, including activity traces from self-monitoring wearables, and use them to develop models that are able to anticipate disease such that preventive interventions can still be effective. Our hypothesis is that signals for detecting the early onset of chronic diseases can be found by appropriately combining multiple kinds of biomarkers. The UK Biobank [3] is our initial and key data resource, consisting of a cohort of 500,000 individuals characterised by genotype, phenotype and, for 100,000 individuals, also free-living accelerometer activity data. Our initial focus is on using UK Biobank data to learn models to predict Type-2 diabetes which is typically associated with insufficient activity and sleep and are preventable given lifestyle changes. [4] Our initial investigation has focused on extracting high-level features from activity traces, and learning traditional models (random forests and SVM) to predict simple disease state. This has given us a baseline for the more complex studies that include genotyping data (selected SNPs), and will be based on deep learning techniques. References [1] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," Nat. Rev. Clin. Oncol., vol. 8, no. 3, p. 184, 2011. [2] N. D . Price et al. , " A wellness study of 108 individuals using personal, dense, dynamic data clouds," Nat. Biotechnol., vol. 35, p. 747, Jul. 2017 [3] C. Bycroft et al., "The UK Biobank resource with deep phenotyping and genomic data," Nature, vol. 562, no. 7726, pp. 203–209, 2018. [4] Cassidy S, Chau JY, Catt M, et al "Cross-sectional study of diet, physical activity, television viewing and sleep duration in 233 110 adults from the UK Biobank; the behavioural phenotype of cardiovascular disease and type 2 diabetes" BMJ Open 2016

26. **Yazan Qarout,** Yordan Raykov, Max Little

Human Behaviour Analysis Through Probabilistic Modelling of GPS data

In urban city planning, it is becoming increasingly difficult to improve and maintain the inhabitants' quality of life and security due to the rising number of the population. Human behavioural characteristics and movement understanding can be an important tool to help assure improvement in the realm of urban planning. However, particularly when examining automated geolocated data (for example GPS data), studies in the field of human movement behaviour analysis are uncommon and often require labels that are difficult to find in real world applications. This is possibly due to the challenges associated with mining and analysing GPS data. It is dynamic, highly-irregularly sampled, noisy, and its data collection can be frequently interrupted due to environmental factors causing missing data. Moreover, different time series trajectories representing different journeys often vary with the number of observations adding challenges to the problem of comparing, clustering and grouping the data sequences. Nevertheless, the lower sensing modality of GPS data has less ethical and privacy concerns to other behaviour monitoring sensors such as CCTV cameras, yet can hold very rich information on behavioural characteristics making it an attractive data source to study. In order to understand the data, compare between multiple trajectories, and bypass the previously mentioned challenges we opted to design a generative probabilistic model for GPS data summarising the high dimensional time series information with the model's parameters and states. Human movement patterns are complex making them difficult to describe with a single set of parameters. Therefore, the generative model structure may better resemble the mechanics of a switching model with parameters optimised for each state in the data sequence corresponding to different behaviour patterns. We propose a novel Non-Homogeneous Vector Autoregressive infinite Hidden Markov (NH-VAR-iHMM) model. Building on the theory of the VAR-HMM structure, we introduce a new discrete and independent semi-markov variable which acts as a parent to the state indicator variable in the Probabilistic Graphical Model (PGM) to represent environmental factors that influence human movement behaviour. The model was applied on 1 weeks' worth of GPS data from taxis in the city of Beijing, China. The results identified the journey points into respective clusters that describe specific behaviour patterns in urban city vehicular travel including peak time, off-peak time, motorway travel and airport journeys.

27. **Yordan Raykov,** Luc Evers, Marjan Farber, Max Little

Probabilistic modelling of gait for remote passive monitoring applications

Passive and non-obtrusive health monitoring using wearables can potentially bring new insights into the user's health status throughout the day and may support clinical diagnosis and treatment. However, identifying segments of free-living data that sufficiently reflect the user's health is challenging. In this work we have studied the problem of modelling real-life gait which is a very indicative behaviour for multiple movement disorders including Parkinson's disease (PD). We have developed a probabilistic framework for unsupervised analysis of the

gait, clustering it into different types, which can be used to evaluate gait abnormalities occurring in daily life. Using a unique dataset which contains sensor and video recordings of people with and without PD in their own living environment, we show that our model driven approach achieves high accuracy gait detection and can capture clinical improvement after medication intake.

28. **Alexia Sampri,** Nophar Geifman, Philip Couch, Niels Peek

Challenges in the aggregation of biomedical datasets and probabilistic approaches to overcome representational heterogeneity

The volume of patient data has grown in an autonomous way with each hospital, electronic device, clinical trial and practicing doctor generating diverse data that are stored in different databases. Putting data together from different sources into a homogeneous data resource would enable unprecedented opportunities to study human health. However, these disparate collections of data are inevitably heterogeneous and have made aggregation a difficult and arduous challenge. In this paper, we focus on the issue of representational heterogeneity: although data stored at different sites may have identical real-world semantics, the data representations methods may differ. In this paper, we argue that there is no need for perfect data standardisation. Based on an example of three datasets systemic Lupus Erythematosus, we suggest the development of an advanced probabilistic methodology that quantifies the uncertainty generated by incomplete harmonisation.

29. **Basabdatta Sen-Bhattacharya,** Sarvesh Kakodkar

Experimenting with speech recognition on the SpiNNaker Machine: a work in progress

Speech recognition is a sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. Here, we report an ongoing work with an objective to implement an existing Natural Language Processer (NLP) and classifier for spoken digits by Maida et al [1]; the implementation is being designed on the SpiNNaker machine, a novel massively-parallel computer architecture, that is built using low-power ARM processors [2]. The algorithm in [1] makes use of a non-recurrent Spiking Neural Network (SNN) that learns to convert speech signal into spike train signature. These signatures are distinguishable from other speech signals representing different words, enabling classification in devices that use only spiking neurons. The non-recurrent SNN implements populations of Izhekevich's neuron models [3]. This is currently implemented on sPyNNaker [4], the underlying toolchain for SpiNNaker. Supervised Learning of synapses in the network are implemented using Hebbian Spike Time Dependent Plasticity [5]. Our implementation of Maida et al's model on SpiNNaker will use the Free Spoken Digit Dataset [6], consisting of a labelled dataset of audio spoken digits. This novel and simple SNN model can be used for further applications like Speech Translation, where the output signatures can be fed to train a Reservoir network for English to Hindi (and vice-versa) speech translation. Reservoir Computing is a framework that maps an input to a higher dimension. A simple readout is trained to read the state of the reservoir. Main benefit is that the training is performed only at the readout and the reservoir is fixed. Such a simple SNN model running on a

low-power device could vastly help the rural Indian education and healthcare systems.

30. **Cameron Shand,** Richard Allmendinger, Julia Handl, Andrew Webb, John Keane

Evolving Controllably Difficult Datasets for Clustering

Synthetic datasets play an important role in evaluating clustering algorithms, as they can help shed light on consistent biases, strengths, and weaknesses of particular techniques, thereby supporting sound conclusions. Despite this, there is a surprisingly small set of established clustering benchmark data, and many of these are currently handcrafted. Even then, their difficulty is typically not quantified or considered, limiting the ability to interpret algorithmic performance on these datasets. We have created a new data generator that uses an evolutionary algorithm to evolve cluster structure of a synthetic data set. We demonstrate how such an approach can be used to produce datasets of a pre-specified difficulty, to trade off different aspects of problem difficulty, and how these interventions directly translate into changes in the clustering performance of established algorithms.

31. **Karin Verspoor,** Dat Quoc Nguyen, Blanca Gallego Luxan

Risk prediction using electronic health records of patients with atrial fibrillation

Electronic health records (EHRs) of patient information, collected during each clinical encounter, have been increasingly available in the last decade. This provides a basis for boosting the quality of healthcare analytics. Structured patient information such as demographics (e.g. age and gender), diagnoses, procedures and prescriptions is extremely useful and can be utilized as machine learning features for predicting stroke, bleeding and mortality risks in patients. This study focuses on predicting health risks in patients with atrial fibrillation (AF) treated with Warfarin from a primary care database (i.e., CPRD—the UK Clinical Practice Research Datalink). We find that a LSTM-based model produces a higher F1 score than those of conventional models SVM and Naive Bayes.

32. **Wenjuan Wang,** Niels Peek, Vasa Curcin, Abdel Douiri, Harry Hemingway, Alex Hoffman, Anthony Rudd, Charles Wolfe, Benjamin Bray

A systematic review of machine learning models for predicting outcomes of stroke

Background In stroke care, there have been interests in the use of machine learning to provide more accurate predictions of outcomes. Due to the complex nature of healthcare data, there are many possibilities for how machine learning can be applied. The aim of this systematic review is to identify these kinds of applications in stroke care, in order to describe the opportunities and challenges of machine learning applications in this field. Objective To carry out a systematic review of published research on applications of machine learning to predict clinical outcomes of stroke. We aimed to summarise which machine learning methods have been used and what their performance was; identify use cases; critically appraise model development and validation methods; and compare the performance of machine learning methods with statistical modelling methods. We focused exclusively on the use of structured clinical data, thus excluding applications of image and text analysis. Methods We carried out systematic searches (Search terms: stroke, machine learning, clinical prediction models and related

synonyms and abbreviations) in Medline and Web of Science. We used previously published search filters where appropriate. Titles and abstracts were screened, and full copies of the included papers were downloaded for review. Results The review is currently in progress and has identified 99 papers from Pubmed and 54 papers from Web of Science. The final results will be ready for presentation at the conference. Discussion We will discuss the key findings of the review and discuss the implications for future research and clinical application of machine learning in stroke care.

33. **Wil Ward,** Mauricio Alvarez

Variational Bridge Constructs for State-Space Gaussian Processes with Non-Gaussian Likelihood

State-space Gaussian processes are a reinterpretation of a Gaussian process as a white-noise driven stochastic differential equation with drift and diffusion terms. As with batch GPs that have Gaussian likelihood, the posterior solution of the SDE can also be inferred exactly. Existing approaches for GP regression in settings with non-Gaussian likelihood include using techniques such as the Laplace approximation or variational Bayes, which have recently been adapted for use with state-space GPs. In this work, we have utilised a variational approach to constructing Brownian bridges to approximate GPs using state-space dynamics. Using black-box variational inference, we can construct an unbiased evidence lower bound and optimise with respect to neural network weights to optimise our estimate of the posterior. Using this black-box approach, we can make use of autodifferentiation, allowing the construction of an estimator for any likelihood without making additional approximation assumptions. Changing the likelihood requires only little tweaking of the underlying model and empirical results demonstrate the flexibility of such an approach.

34. **Andrew Webb,** Charles Reynolds, Dan-Andrei Iliescu, Henry Reeve, Gavin Brown

Joint Training of Neural Network Ensembles

We examine the practice of joint training for neural network ensembles, in which a multi-branch architecture is trained via single loss. This approach has recently gained traction, with claims of greater accuracy per parameter along with increased parallelism. We introduce a family of novel loss functions generalizing multiple previously proposed approaches, with which we study theoretical and empirical properties of joint training. These losses interpolate smoothly between independent and joint training of predictors, demonstrating that joint training has several disadvantages not observed in prior work. However, with appropriate regularization via our proposed loss, the method shows new promise in resource limited scenarios and fault-tolerant systems, e.g., IoT and edge devices. Finally, we discuss how these results may have implications for general multi-branch architectures such as ResNeXt and Inception.