Data Science Campus

MDataGov Symposium 2020

Master's in Data Analytics for Government Symposium 2020

The Studio, Birmingham, B2 5EP

25th-26th March 2020

Data Science Campus
MDataGov Symposium 2020

WELCOME TO THE MDATAGOV SYMPOSIUM 2020!

We are delighted to welcome you to the 3rd annual Master's in Data Analytics for Government Symposium.

We have designed this event for former, current and prospective students to come together to showcase their exciting data science research and projects and to share experiences and best practice.

We have a packed timetable for the next two days which will include:

- Presentations from MDataGov university providers
- Exciting presentations and workshop from fellow students on a wide range of data science topics
- Plenty of networking opportunities!

We would like to thank the ONS Data Science Campus, university providers, speakers and participants for supporting this event.

We hope that you enjoy the conference.

See you in Birmingham!

The 2020 MDataGov Symposium Organising Committee:

Khloe Evans, Office for National Statistics

Wei (Tony) Guo, HM Revenue and Customs

Lucy Gwilliam, Office for National Statistics

Sarah Liley, NHS Digital

Elizabeth Scopel, Office for National Statistics

Rajni Sandhu, Office for National Statistics

Solange Correa-Onel, Office for National Statistics

Data Science Campus
MDataGov Symposium 2020

PROGRAMME

DAY 1 – Wednesday, 25th March 2020, Room Innovate

13:00 – 13:30	**Registrations and refreshments**
	Session 1 – Chair: Lucy Gwilliam, ONS
13:30 – 13:50	Welcome and Overview, Solange Correa-Onel, ONS Data Science Campus
13:50 – 14:50	Bayesian methods and data analytics for government, Dr Jason Hilton, University of Southampton
14:50 – 15:00	**Break**
	Session 2 – Chair: Sarah Liley, NHS Digital
15:00 – 16:00	AI ethics and ethical AI in the age of machine learning Dr Matthias Rolf, Oxford Brookes University
16:00 – 17:00	Dimension reduction - visualisation and analysis of high-dimensional data Dr Tengyao Wang, University College London
17:00 – 19:00	**Networking** **(speed mentoring sessions available - please sign up with Liz Scopel on the day)**

Note: 1. Speakers will have 15 minutes to present their work and 5 minutes for questions and answers. 2. It is anticipated that an article with Symposium highlights will be published on the ONS Data Science Campus website (https://datasciencecampus.ons.gov.uk/) shortly after the event. The article will include the Symposium programme and links to the presentations. Please ensure your presentations only include information that can be made available to the public.

Data Science Campus
MDataGov Symposium 2020

DAY 2 (Morning) – Thursday, 26th March 2020, Room Innovate

09:00 - 09:50	**Registrations and refreshments**
09:50 - 10:00	Welcome, the Organising Committee
10:00 - 10:55	**Session 3 – Chair: Rajni Sandhu, ONS**
	Keynote Speaker *Introducing Data Science to International Development* Tom Wilkinson, Data Science Hub Site Lead (East Kilbride) and Head of Data Science, Department for International Development
	Session 4 - Chair: Khloe Evans, ONS
11:00 - 11:20	Presentation 1: *Seeing trees in a forest: modelling, aggregating and individuals* Dean Gordon, Northern Ireland Statistics and Research Agency
11:20 - 11:40	Presentation 2: *Robust text analytics using information retrieval tools: using traditional Information Retrieval (IR) algorithms and tools to query and summarize a corpus* Martin Wood, Office for National Statistics
11:40 - 12:00	Presentation 3: *Cross-government collaborative working to improve the coherence of adult social care finance data* Sarah Liley, NHS Digital
12:00 - 13:00	**Lunch**

Data Science Campus
MDataGov Symposium 2020

DAY 2 (Afternoon) – Thursday, 26th March 2020, Room Innovate

Session 5 - Chair: Wei (Tony) Guo, HMRC

13:00 - 13:20 Presentation 4:
The application of computer vision and machine learning linking to data science
Li Chen, Office for National Statistics

13:20 - 13:40 Presentation 5 (joint talk):
Becoming a Data Analyst, Lucy Gwilliam, Office for National Statistics
Insights from a first year MDataGov student, Ingrid Bukirwa, HM Revenue and Customs

13:40 - 14:00 Presentation 6:
Learning to predict unseen views
Edward Bartrum, PhD student, University College London and Alan Turing Institute

14:00 - 14:20 Presentation 7:
Implementation of big data technologies in the Civil Aviation Authority
Panagiota (Giota) Pantazopoulou, Civil Aviation Authority

14:20 - 14:40 **Coffee Break**

14:40 - 15:40 **Session 6 - Workshop on "Managing a data science project"**
Led by: Rajni Sandhu (ONS), Sarah Liley (NHS Digital), Jazz Grimsley (ONS) and Isabela Breton (ONS)

Session 7 - Chair: Lucy Gwilliam, ONS

15:40 - 16:00 Presentation 8:
Assessing data science capability across the public sector
Harrison Davies, Office for National Statistics

16:00 - 16:05 **Best Presentation voting,** Wei (Tony) Guo, HMRC

16:05 - 16:10 **Winner announcement and closing**, David Johnson, ONS Data Science Campus

Note: 1. Speakers will have 15 minutes to present their work and 5 minutes for questions and answers. 2. It is anticipated that an article with Symposium highlights will be published on the ONS Data Science Campus website (https://datasciencecampus.ons.gov.uk/) shortly after the event. The article will include the Symposium programme and links to the presentations. Please ensure your presentations only include information that can be made available to the public.

ABSTRACTS

MDataGov University Providers

Bayesian Methods and Data Analytics for Government

Dr Jason Hilton, University of Southampton

Abstract: We live in an uncertain world. Robust decision-making should take into account our lack of certainty about both the present and the future. Bayesian statistics provides an intuitive approach to the quantification of uncertainty that allows for the incorporation of prior information about the system of interest. This session will provide a brief introduction to Bayesian methods, and will give a simple demonstration of the use of the `Stan` software package for estimating Bayesian models of population change. The Stan software allows the specification of generative models in an accessible probabilistic programming language, and employs Hamiltonian Monte Carlo techniques to efficiently obtain posterior samples from this model. Methods for checking the validity of the model and the convergence of the sample will also be demonstrated using Stan's interface with the R programming language.

AI Ethics and Ethical AI in the age of machine learning

Dr Matthias Rolf, Oxford Brookes University

Abstract: The advance of machine learning techniques and readily available data creates many opportunities to advance businesses and public services, but also poses serious ethical challenges. In this talk I will outline common practical ethical challenges and purposes with learning machines as tackled at Oxford Brookes University under the umbrella of the Institute for Ethical AI (IfEAI). Example projects include the application of ML in contract law analysis, and visitor flow prediction at one of UK's most iconic heritage sites. I will further give perspectives how continuously learning machines may in the future be able to fully absorb ethical and social values through social embedding and current developments in reinforcement learning. Robotics examples will serve to illustrate the need for such models of morality.

Dimension reduction -- visualisation and analysis of high-dimensional data

Dr Tengyao Wang, University College London

Abstract: In many modern Big Data settings, data sets routinely exhibit high dimensionality, in the sense that the number of features measured can be of comparable or even larger order than the number of observations. High-dimensional data have posed new methodological challenges to data scientists and practitioners. Dimension reduction techniques have been developed as a response to these challenges. In this talk, we will look at a few different dimension reduction techniques, and their applications in both data visualisation and subsequent statistical analysis. This includes Principal Component Analysis and its sparse variant (Sparse PCA), kernel variant (kernel PCA), nonlinear dimension reduction techniques (e.g. tSNE and UMAP) and applications in genomics and image analysis.

Keynote Speaker

Introducing Data Science to International Development

Tom Wilkinson, Data Science Campus hub site lead (East Kilbride) and Head of Data Science, Department for International Development

Abstract: In this presentation, Tom Wilkinson will talk about his own career experience across government as well as those for the first few data scientists in the Department for International Development (DFID). Tom originally joined government hoping to work for DFID, and after three years as a Fast Stream Operational Researcher in applied Maths and data science roles across justice and security, Tom took advantage of a new appetite in DFID to exploit data to achieve his goal. Along the way he found that the Civil Service can be flexible to individual interests and was very receptive to people taking the initiative and volunteering for whatever interested them outside their main job. DFID's first few data scientists have generally enjoyed similarly quick progression, and flexibility to follow their intellectual interests. Tom will try to sum up the behaviours that get the most out of these opportunities in the Civil Service, while giving a tour of some past and future applications of data science in international development, from coaching others through dashboarding and automation to applying various data science techniques to answer business and development questions.

Contributed Sessions

Presentation 1

Seeing trees in a forest: modelling, aggregating and individuals, Dean Gordon, Northern Ireland Statistics and Research Agency

Abstract: This presentation looks firstly at attempts made to model pupil choices for post-primary schools. It then discusses the difficulty in making decisions based on such models and looks at a simpler, but fuller, exploration of the data and discusses potential improvements to the data sources that would reduce the need for models. The presentation concludes with a discussion of mapping, principally as a tool to allow investigations to drill down to individual data points, but also to look at spatial relationships.

Presentation 2

Robust text analytics using information retrieval tools: using traditional Information Retrieval (IR) algorithms and tools to query and summarize a corpus, Martin Wood, Office for National Statistics

Abstract: Topic models and Word2Vec-style semantic/vector representations of text often don't work brilliantly with "middling"-sized data, and tracking the evolution of topics through time is an area of active research. Additionally, topics != things; in many contexts, especially current events, the researcher is more interested in identifying specific entities or phrases of interest. Here are presented alternative statistical tools from the field of Information Retrieval (IR) that can be used to identify and track documents and subjects of interest within text corpuses, using freely available open source tools that are easy and rapid to set up. Using a corpus of recent ("middling"-sized) news article snippets, ElasticSearch is used to demonstrate retrieval and relevance scoring, using stemming and fuzzy matching to make the corpus robustly searchable. Visualisations are built within Kibana's dashboards to power text summarisation, and these are enhanced using some additional IR-based pre-processing steps of phrase detection and POS tagging that can be implemented in Python. Text data timeseries are used to identify recently emerging entities of interest. The methods and tools presented allow for rapid development of robust text analytics and horizon scanning capabilities.

Presentation 3

Cross-government collaborative working to improve the coherence of adult social care finance data, Sarah Liley, NHS Digital

Abstract: Over £17bn of public money is spent on adult social care each year: it is no surprise that decisions over how that money is spent attracts scrutiny and, as a statistician, I have to ensure that there is a common understanding of the data I am responsible for. We formed a cross-government working group, with representation from DHSC, MHCLG, NHS Digital and NHS England, to align the national and local reporting of Adult Social Care (ASC) expenditure, including the reconciliation of two expenditure datasets published by NHS Digital and MHCLG. With funding for social care so often in the headlines, it is important that everyone is clear about current levels of expenditure, where funding comes from and the data sources. In this talk, I will present the quantitative and qualitative analysis undertaken in reconciling two National Statistics collections produced by two different organisations, and the benefits and lessons learned from the experience of coordination across four large organisations. Adapted from "How cross-government collaborative working is informing the Adult Social Care policy debate", presented by DHSC, MHCLG and NHS Digital in a parallel session at the 2019 GSS Conference.

Presentation 4

The application of computer vision and machine learning linking to data science, Li Chen, Office for National Statistics

Abstract: I will present some application cases such as document imaging processing and video compression: How to apply advanced techniques from computer vision and machine learning in the industry of information service and broadcasting. I will talk about how these techniques can benefit to data science – which help me get job offer from Data Science Campus. I will also share my experience on the methodology changes and trends through these years in industry due to evolutions of computing environment.

Presentation 5 (joint talk)

Becoming a Data Analyst, Lucy Gwilliam, Office for National Statistics

Abstract: After finishing a Mathematics BSc Degree at Aberystwyth University, I joined the ONS Data Science Campus (DSC) as a Data Analytics Apprentice. The structure of the apprenticeship was a year in DSC and then two six-month placements in business areas within ONS. During my time in DSC, I worked on some very exciting projects which used innovative data science techniques such as machine learning and use of application programming interfaces (APIs). My first year as an apprentice allowed me to get a good working knowledge of data analytics and data science. After my year in the Campus I moved to the Sustainable Development Goals (SDG) team, where I have had a chance to use my data analytics skills on a number of projects. For example, using Python to automate data collection using the ONS API and to automate data processing tasks. After finishing the six-month placement, I stayed on the SDG team as a Data Analyst and not long after I completed the apprenticeship, I applied to do the MSc Data Analytics for Government to expand my skills and knowledge at a higher level. In this talk, I will give an overview of my experience as data analyst and present some of the research projects I have been involved with.

And

Insights from a first year MDataGov student, Ingrid Bukirwa, HM Revenue and Customs

Abstract: I am currently a statistician at HMRC and was fortunate to join the MDataGov programme in September 2019 at Brookes as a Campus-funded student. I wanted to challenge myself, build on my knowledge and have a more confident understanding of mathematics and statistics to connect the dots between the two subjects with data science – to broaden my understanding and application of my skillset. I've completed Data foundations, Regression Modelling and Statistical Programming modules and I'm currently taking Introduction to Machine Learning (at ONS) and Statistics for Government at Oxford Brookes. In this presentation, I will talk about my experience so far and how the MDataGov programme is equipping me with essential skills to meet business need and confidently bring new ideas to improve my work at HMRC.

Presentation 6

Learning to predict unseen views, Edward Bartrum, PhD student, University College London and Alan Turing Institute

Abstract: Since the introduction of GANs, there has been rapid progress in the direction of realistic image synthesis, and more recently there has been increased interest in controllable image generation. Some recent works allow synthesis to be guided by a segmentation map or user sketch whilst others develop models which admit geometric transformations on generated content. On a related track there has been progress in differentiable rendering, seeking to extract 3D geometry and texture information from a 2D view of an object by optimising a rendering process with approximate gradients. This talk will discuss research at the intersection of these tracks, aiming to realistically render objects into new poses under arbitrary mesh transformations, based on a single input image.

Presentation 7

Implementation of big data technologies in the Civil Aviation Authority, Panagiota (Giota) Pantazopoulou, Civil Aviation Authority

Abstract: An overview of the Civil Aviation Authority (CAA)'s journey to date, to implement big data technologies in order to extract insights and actionable intelligence from existing data sources and to facilitate the accommodation of new data sources to enrich the existing knowledge base will be provided. We will look at the main challenges (technical and cultural) and the skills and competences required to implement and endorse new tools and concepts to shift current Excel dominated environments and analyses to platforms that favour automation and self service. Examples and practical applications of data analysis and visualisations will be shared to demonstrate the value and the benefits of moving towards data science techniques.

Presentation 8

Assessing data science capability across the public sector, Harrison Davies, Office for National Statistics

Abstract: The Government Data Science Partnership (GDSP) has undertaken an audit of data science capability across the UK public sector. Commissioned on behalf of HM Treasury, the audit will help to ensure the public sector and its employees realise the maximum potential of data. This presentation will provide: i) an overview of the audit; ii) findings on the data science skill level of analysts across government and ii) the introduction of a product that enables the self-assessment of organisational data science capability. It will also include the actions presented to HM Treasury. This presentation will conclude with the next steps of enabling the Data Science Campus, GDS Academy and the Government Statistical Service to develop and deliver a data science capacity-building programme.

Workshop: *Managing a data science project,* led by: Rajni Sandhu (ONS), Sarah Liley (NHS Digital), Jazz Grimsley (ONS) and Isabela Breton (ONS)

Description: This is an interactive workshop exploring the development of data science projects. The role of the data scientist will be explored- specifically how to be agile in their approach to developing projects. Several types of projects tackled by data scientists will be introduced including; predictive, descriptive, experimental. Participants will explore how the same question can be addressed using these different project types.