

**Government Statistical Service Conference 2017**

Sheffield United Football Club

22 – 23 November 2017

Pioneers: on the forefront of statistics and data science



Sponsored by:

#GSSConf

Foreword 1 Venue 2 Agenda 3 Keynote Speakers 7 Parallel Sessions 12 Stalls 32 Committee 34

**Foreword**

**John Pullinger, National Statistician**

Three years ago the Government Statistical Service embarked on a five year strategy: Better Statistics, Better Decisions. The strategy set an expectation that UK official statistics would be radically different, mobilising the power of data to help Britain make better decisions. It set out a call to action for each person working in the statistical system to think about their own role and ask: how can I be more helpful to those I serve; how can I demonstrate professionalism by delivering a high quality service that will be trusted; how can I innovate and make things better; how can I be more efficient in my use of resources, money and time; and how can I develop my capability and learn new things?

As I look back, I have seen many examples where the possibilities arising from previously unimaginable sources of data have been grasped. I have seen true commitment to our values of honesty, integrity, objectivity and impartiality - respecting the interests of those whose data are being utilised for public benefit and by never straying from description into advocacy.

Looking ahead I am excited about all the possibilities that we can and must grasp. This year's conference is about being pioneers. That is a spirit we can all take away from the conference. I hope this year inspires you all to look at statistics in a new way, a way where the benefits are realised by all.

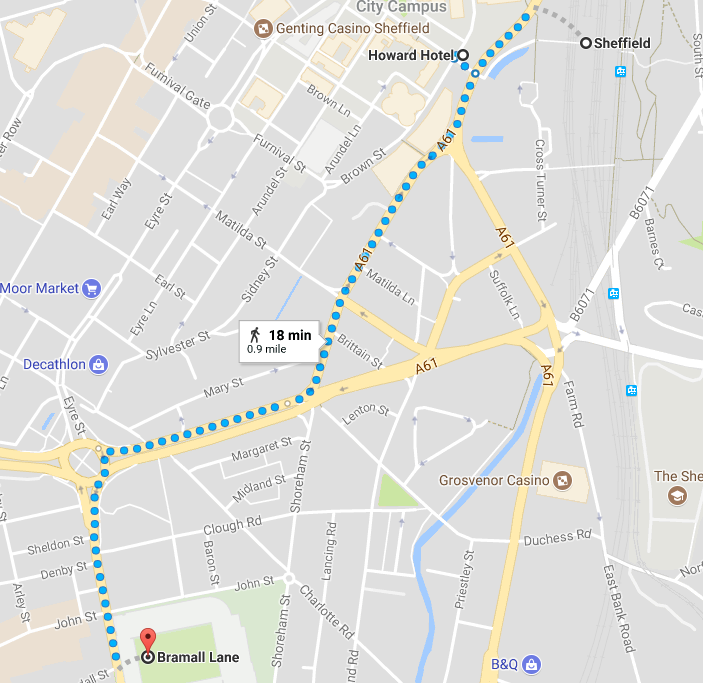
**Alex Miller, Conference Committee Chair**

I’d like to welcome you all to the 2017 GSS conference in my home city of Sheffield. Whilst I have had the honour of chairing the organising committee, the committee of volunteers have done all the serious work to put on this fantastic event.

This is a wonderful opportunity to hear about some exciting developments both from within the GSS and the wider statistical community. We have talks on the analytical work following the Grenfell Tower fire, and a talk about how data can be used to identify potential victims of human trafficking. These are very potent examples of how our work can have a huge impact on people’s lives.

The conference is also an important chance to strengthen the GSS as a network. Whatever statistical issues you are facing at work, someone else in the GSS has probably addressed the same issue in a different context. That is why I think that networking at an event like this is so valuable. So please take time to talk to each other. The person sat next to you at dinner might be the person who will help you solve that tricky problem next week.

I hope you all enjoy the conference, take new ideas back to your departments and get the chance to catch up with old colleagues and meet new ones.



**Conference** **Venue**

**Train Station**

**Bus Stop**

**Venue Details**

Conference venue: Sheffield United Football Club

Bramall Lane, Sheffield, S2 4SU

The venue is a 15-20 minute walk from Sheffield Train Station. Exit the station and turn left onto Sheaf Street. Follow the road right as it joins St Mary’s Road. Take the first left at the roundabout onto Bramall Lane.

Alternatively, Howard Hotel bus stop is located just across Sheaf Square; the number 18, 19 and 252 services stop at the venue. Taxis are available outside the station.

|  |  |  |  |
| --- | --- | --- | --- |
| **DAY 1 –** **22nd November 2017** | | | |
| 10:00 – 11:30 | Platinum Suite | Registration (luggage store, refreshments and exhibition stalls) | |
| **Keynote Session 1** | | | |
| 11:30 – 13:00 | Platinum Suite | **Welcome from the National Statistician** (John Pullinger)  **Code of Practice** (Ed Humpherson – Director General of Statistical Regulation)  **Frontiers of data science and statistics** (Tom Smith - Managing Director of ONS Data Science Campus and Tracey Brown - Director of Sense about Science) | |
| 13:00 – 14:00 | Platinum Suite | **Lunch/Exhibition Stalls** | |
| **Parallel Session 1** | | | |
| 14:00 – 15:00 | **1.1**  Legends of the Lane | Reproducible analytical pipelines | Matthew Upson et al (Cabinet office) |
| **1.2**  Directors Boardroom | Development of a Data Science Lab to save £1 billion – 3 years in  Fraud and error measurement: Adapting to a changing benefit landscape | Nadine Morrisroe  (NHS BSA)  Jennifer Snape (DWP/OME) |
| **1.3**  Platinum Suite | Understanding the potential - creating a cross-GSS data family: In a huge statistical system how can we solve the problem of finding the right data first time?  Making health and care statistics accessible | Darren Barnes (ONS)  and Sam Hall (ONS)  Helen Colvin (ONS) |
| **1.4**  1889 Suite | Data driven school places, Uganda, & R Shiny | Sam Cuthbertson (DFE) |
| **1.5**  International Bar | How to build data science capability and not go mad in the process  Data science at MoJ | Tom Ewing (DFT)  Jon Roberts (MoJ) |
| 15:00 - 15:30 | Platinum Suite | **Break/Exhibition Stalls** | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Parallel Session 2** | | | |
| 15:30-16:30 | **2.1**  1889 Suite | Using data science to transform research and innovation policy  Combining information in multiple spatial datasets to produce estimates and standard errors of population areas | Juan Mateos-Garcia (NESTA)  Alan Brewer (National Forest Inventory) |
| **2.2**  Platinum Suite | Machine learning in production: transforming visa casework and identifying potential victims of trafficking  Exploring mental wellbeing from prisoner case-notes using text mining | Tom Wakeford (HO)  Jo Lee (MoJ) |
| **2.3**  International Bar | Integrated Data Enabling Analysis and Statistics (IDEAS): A new data model for integrating data  Household statistics from a future administrative data census | Becky Tinsley (ONS)  Peter Jones (ONS) |
| **2.4**  Directors Boardroom | Refreshing the Code | Penny Babb (ONS) |
| **2.5**  Legends of the Lane | GSS Good Practice Team: capability building  Random Forest Analysis: the prediction of fragile A&E systems | Martin Ralphs (ONS)  Ed Beake (DH) |
| **Keynote Session 2** | | | |
| 16.30 – 17:30 | Platinum Suite | **Making Sense of Big Data**  Maria Fasli - Director of Institute for Analytics and Data Science/ UNESCO Chair in Analytics and Data Science  Daniel Hulme - Founder and CEO of Satalia. Researcher/teacher in Information Security, Technology Entrepreneurship and Information Studies at UCL. Visiting Fellow of the Big Innovation Centre. | |
| 19:30 – onwards | Platinum Suite | **Pre-Dinner Welcome from the National Statistician** (John Pullinger)  **Conference Dinner** (20.15 onwards) | |

|  |  |  |  |
| --- | --- | --- | --- |
| **DAY 2 –** **23rd November 2017** | | | |
| 9:00-9:20 | Platinum Suite | Refreshments and exhibition stalls | |
| **Keynote Session 3** | | | |
| 09:20 – 10:30 | Platinum Suite | **Panel discussion: Statistical Literacy**  **Opening discussion:** David Spiegelhalter (President of the Royal Statistical Society)  John Pullinger (National Statistician)  Sarah Henry (Director of Methods, Data and Research at ONS)  Gurleen Popli (Economics Lecturer at the University of Sheffield)  Chair: Martin Ralphs (Head of GSS Good Practice Team) | |
| **Parallel Session 3** | | | |
| 10:30– 11:30 | **3.1**  Legends of the Lane | Statistical modelling of arrears history data for decision support in  Social Housing Management  Race Disparity Audit: Building a data website and collaborating across Government. | Farida Mustafazade (Newcastle University)  Zamila Bunglawala (Cabinet Office) |
| **3.2**  Directors Boardroom | Quota sampling guidance  Quality Assurance of Administrative Data (QAAD): The population statistics perspective | Gary Brown (ONS) and Catherine Mottram (DfT)  Philip Humby (ONS) |
| **3.3**  International Bar | The ‘Data’ Science of Where - the role of Geospatial analysis in Data Science and Statistics  Disaggregating statistics using spatial techniques | Steve Croney (ESRI)  Chris Gale (ONS) |
| **3.4**  Platinum Suite | DCLG’s analytical work in the immediate aftermath of the Grenfell Tower fire | Sandra Tudor (DCLG) |
| **3.5**  1889 Suite | Data ethics for statistics and research  Better Scraping, Better Statistics? Using web-scraped data in statistical outputs | Petros Saravakos (UKSA)  Matthew Greenaway (ONS) |
| 11:30 - 11:45 | Platinum Suite | **Break/Exhibition Stalls** | |
| **Parallel Session 4** | | | |
| 11:45 – 12:45 | **4.1**  International Bar | Improving data communication, 90+ tables to an API  The use of Data Science to improve the way we process and disseminate data | Hiren Bhimjiyani (BEIS)  Tom Davies and Mark Burley (DWP) |
| **4.2**  Platinum Suite | The Dissemination Game: How to communicate official statistics to non-expert users  Understanding crime in prisons: the problem | Hannah Thomas (Gov Wales)  Jack Tattersall (MoJ) |
| **4.3**  Legends of the Lane | Cracking address matching: Why it matters & why it’s not easy  Use of the VAT data within ONS, in collaboration with the ONS Data Science Campus | Alistair Calder (ONS)  Andrew Sutton (ONS) |
| **4.4**  1889 Suite | Global GSS: International priorities, preparing for Brexit, and International Development | Emily Poskett (ONS), Robert Bumpstead (UKSA), Will Laffan (UKSA) and Ian Cope (ONS) |
| **4.5**  Directors Boardroom | Discover how design of local policy is elevated by engaging data analysis: a LEP perspective  Why should we have a ‘place-based’ industrial strategy anyway? | Lisa Clark (Sheffield City Region Executive Team)  Paul Swinney (Centre for Cities) |
| **Keynote Session 4** | | | |
| 12:45 – 13:30 | Platinum Suite | **Close: Statistics for the public good**    Claire Melamed - Director of Global Partnerships for Sustainable Development Data | |
|  | | | |

**Keynote Speakers**

**John Pullinger**

John Pullinger started as the National Statistician, Head of the Government Statistical Service (GSS) and Chief Executive of the UK Statistics Authority on 1 July 2014. He was previously Chair, and Vice-Chair, of the United Nations Statistical Commission (UNSC). In this capacity John was greatly involved in the development of measures for the Sustainable Development Goals.

John's career began in 1980 when he joined the civil service after graduating in geography and statistics from Exeter University. After several statistical roles in different departments, John joined the Central Statistical Office as a senior civil servant in 1992. He was the project manager for the creation of the Office for National Statistics and was the policy lead on the development of the GSS. He has worked on diverse projects and topics, including responsibility for flagship publications like 'Social Trends', leading the neighbourhood statistics programme as well as being actively involved in the creation of the Statistics Commission and National Statistics in 2000.

In 2004, John became the 14th librarian to the House of Commons. He continued to be involved in the statistical community and took up the presidency of the Royal Statistical Society on 1 January 2013.

John is married with three adult children and three grandchildren. He is active in his local community in Tunbridge Wells and has been the chairman of Great Culverden Park Limited since 1999.

**Ed Humpherson**



Ed Humpherson was appointed as Director General for Regulation in October 2013 and took up post in January 2014.

Prior to joining the Authority, Ed was a Board Member and Executive Leader for Economic Affairs at the National Audit Office, a post he held since July 2009. This role included responsibility for the overall strategic direction of NAO’s work on economic affairs.

Between 2007 and 2009, Ed Humpherson was Assistant Auditor General, and before that was NAO’s Director of Regulation between 2003 and 2007. Ed joined the National Audit Office in 1993.

Ed Humpherson was educated at the University of Edinburgh where he obtained a first-class MA honours degree in Politics and Economic History. Ed is a Chartered Accountant and a member of the Institute of Chartered Accountants in England and Wales. Ed is married and has three children.

**Tom Smith**

Tom is managing director at the UK government's Data Science Campus, which explores how new data sources such as earth observation data, images and social media can help us better understand the economy and society. A data addict with 20 years’ experience using data and analysis to improve public services, Tom originally trained as a physicist with a PhD in training neural networks for robot control. Before joining government, Tom was co-founder and chief executive of OCSI, a research and data 'spin-out' company from the University of Oxford that has worked with 100s of government agencies, including leading the government’s Indices of Deprivation used to allocate more than £1 billion per year. He is vice-chair of the Royal Statistical Society Official Statistics section, previously chair of the Environment Agency Data Advisory Group, and member of the Open Data User Group ministerial advisory group.

**Tracey Brown OBE (Director, Sense about Science)**

Tracey Brown has been the director of Sense about Science since 2002. Under her leadership, the charity has turned the case for sound science and evidence into popular campaigns to urge scientific thinking among the public and the people who answer to them. It has launched important initiatives including AllTrials, a global campaign for the reporting of all clinical trial outcomes; and the Ask for Evidence campaign, which engages the public in requesting evidence for claims. In 2010, the Times named Tracey as one of the ten most influential figures in science policy in Britain and in 2014 she was recognised by the Science Council for her work on evidence-based policy making. In June 2017 Tracey was made an OBE for services to science.

**Prof Maria Fasli**

Maria Fasli is a Professor of Computer Science (Artificial Intelligence) and the Director of the Institute for Analytics and Data Science (IADS) at the University of Essex. She obtained her BSc in Informatics from the Technological Education Institute of Thessaloniki in 1996, and her PhD in Computer Science from the University of Essex in 2000. She has held research and academic positions at the University of Essex since 1999 and became Professor in 2012. In 2009, she became the Head of the School of Computer Science and Electronic Engineering at Essex, a post which she held until the end of 2014. In August 2014, she was appointed in her current role as Director of IADS. In 2016, she was awarded the first UNESCO Chair in Analytics and Data Science.

Her research interests lie in artificial intelligence techniques for analysing and modelling complex systems and structured and unstructured data in various domains. Her research has been funded by National Research Councils in the UK, InnovateUK as well as businesses. Maria has published over 120 papers in the field of artificial intelligence has been involved in organising international research events and conferences. She is also passionate about education and pedagogic innovation and in 2005, she was awarded a National Teaching Fellowship by the Higher Education Academy (UK) for her innovative approaches to education.

**Daniel Hulme**

Daniel is the CEO of Satalia, a company that provides Artificial Intelligence (AI) inspired solutions to solve industries hardest problems. He is also the Director of UCL’s Business Analytics MSc, applying AI to solve business and social problems.

Daniel is a popular public speaker specialising in the topics of Philosophy, Technology, Innovation and Organisational Design. He is a serial TEDx speaker and a speaker for the Singularity University.

Daniel has a Masters and Doctorate in AI from UCL, he lectures in Computer Science and Business, and is passionate about how technology can be used to govern organisations and bring positive social impact.

Daniel is the co-founder of ASI Data Science and has Advisory and Executive positions in many companies. He holds an international Kauffman Global Entrepreneur Scholarship and actively promotes positive entrepreneurship and technology innovation across the globe.

**David Spiegelhalter**

David Spiegelhalter is Winton Professor for the Public Understanding of Risk and Fellow of Churchill College at Cambridge University. His background is in medical statistics, particularly the use of Bayesian methods in clinical trials, health technology assessment and drug safety.

As Chair of the Winton Centre for Risk and Evidence Communication, he works to improve the way in which risk and statistical evidence are taught and discussed in society. He gives many presentations to schools and others, advises organisations and government agencies on risk communication, and is a regular commentator on current risk issues. He presented the BBC4 documentaries ‘Tails you Win: the Science of Chance” and the award-winning “Climate Change by Numbers”, and in 2011 came 7th in an episode of Winter Wipeout on BBC1.

He has over 200 refereed publications and is co-author of 6 textbooks, as well as The Norm Chronicles (with Michael Blastland) and Sex by Numbers. He is a Fellow of Churchill College Cambridge, an Honorary Fellow of the Institute for Risk Management, an Honorary Fellow of the Royal College of Physicians, and was elected Fellow of the Royal Society in 2005. He has contributed to many formal investigations, including the inquiries into Bristol children’s heart surgery, Harold Shipman’s murders, and the breast implant scandal.

He was awarded an OBE in 2006 and knighted in 2014, both for services to medical statistics.

He is @d\_spiegel on Twitter, and his home page is

[http://www.statslab.cam.ac.uk/Dept/People/Spiegelhalter/davids.html](http://www.statslab.cam.ac.uk/Dept/People/Spiegelhalter/davids.html%20)

**Sarah Henry**

Sarah joined the Office for National Statistics in 2017 from Manchester City Council and the Greater Manchester Combined Authority where she developed evidence and innovative analysis that informed many local decisions as well as the mayoral devolution deals. Championed innovations include a ward level population forecasting model to enable planning of crucial local services, pioneering linking datasets to inform safeguarding services and early interventions and prevention and a spatial forecasting model to predict ‘urban flow’ in partnership with Ordnance Survey. At ONS she is the Director of Methods, Data and Research and leads the office’s strategy to bring in datasets, prepare them for the production of statistics and make them available for research inside and outside the office.

Sarah is passionate about using data for public good and ensuring that the national data asset is utilized to describe and explain the society and the economy. Data have the power to save lives and fuel the economy but to do so reliably they must be used safely and responsibly, ensuring the analysis is based on sound methods and privacy is protected.

**Claire Melamed**



Dr Claire Melamed is the Executive Director of the Global Partnership for Sustainable Development Data, headquartered in Washington, DC.  She is based in London and was previously a Managing Director at the Overseas Development Institute, has worked for a number of international NGOs, the United Nations, and taught at the University of London and the Open University. She can be found on Twitter at @clairemelamed.

**Gurleen Popli**

Gurleen is a Lecturer in the Department of Economics. She has a B.A. (Economics) from University of Delhi, a M.A. (Economics) from Delhi School of Economics and a Ph.D. (Economics) from University of California, Riverside. She joined the Department of Economics at Sheffield in September 2004, having previously taught at universities in the UK and the USA. Gurleen is Director of Research at the [Institute for Economic Analysis of Decision-making (InstEAD)](http://instead.group.shef.ac.uk/) at the University of Sheffield.



**Parallel Sessions**

|  |  |
| --- | --- |
| **22nd November 2pm-3pm** | |
| **1.1**  Legends of the Lane | **Reproducible analytical pipelines**  **Matthew Upson et al (Cabinet office)**  Producing accurate and timely statistics is a key function of statisticians in government. Statistical publications need to be produced accurately, on time, and reproducibly. At any point in the future, we should be able to reproduce all the steps required to produce a statistic; however manual processes (common in many publications) can make this challenging.  Reproducible Analytical Pipelines (RAP) is a new approach that borrows techniques from software development and reproducible research to reduce the manual burden of statistics production. By using open source technology, the statistics can be produced more quickly, with automated quality control, and in a way that can easily be reproduced, shared and copied. This open and transparent methodology leads to higher quality, more efficient statistics, which better suit our users’ needs.  In this presentation, colleagues from MOJ, DfE, and DCMS, with the help of GDS will take you through a journey of how tools & techniques such as Rmarkdown, unit testing, and continuous integration (commonly employed by software developers), are being implemented for statistics production in their departments. |

|  |  |
| --- | --- |
| **1.2** Directors Boardroom | **Development of a Data Science Lab to save £1 billion – 3 years in**  **Nadine Morrisroe (NHS BSA)**  The NHS Business Services Authority (NHSBSA) is a Special Health Authority and an Arm’s Length Body of the Department of Health which provides a range of critical central services to NHS organisations, NHS contractors, patients and the public. We are data owner of a vast array of transactional data such as NHS prescriptions data, NHS Pensions data and NHS dental data. £34 billion worth of NHS activity flows through our organisation on an annual basis.  The NHSBSA was asked to collaborate to deliver £1 billion of annual recurring savings by March 2018. This presentation describes the creation of a Data Lab and how a team of analysts, data scientists and statisticians have identified over £800m of potential savings within first 3 years and are now working with the wider NHS to action the insight.  It includes:   * explaining how the development of the data lab has involved creating one environment in which we can analyse large volumes of data from our operational systems to derive insight. * how we work with our customers * case studies of projects we have worked on, including analytical methods / models used, key findings and outcomes * the benefits of the data lab and lessons learned along the way.   **Fraud and error measurement: Adapting to a changing benefit landscape**  **Jennifer Snape (DWP/OME)**  The Department of Work and Pensions, Fraud and Error Measurement Analysis Team (FEMA) are responsible for estimating the amount of fraud and error in the benefit system and the expenditure lost to the taxpayer because of this. For example, in the last financial year it is estimated that £3.5billion was overpaid to claimants! Our estimates provide vital insight on such a high profile issue both within the department and to the wider public.    Our team covers all aspects of the measurement process from sample design to data analysis and publication. For a small team, FEMA are facing significant challenges in the years ahead including reductions in the number of performance management visiting staff, pressure to deliver efficiency savings and adapt to a rapidly changing benefit landscape with the roll out of Universal Credit and Personal Independence Payment. These challenges have meant we need to re-think our approaches to measuring fraud and error through new developments and automation. We are constantly evolving as a team, embracing change and innovation, making FEMA an exciting place to be!    This presentation will focus on changes in what we measure (roll out of UC, reduction in legacy benefits, complex methodology etc.), changes in how we measure (introduction of desk based reviews, new sampling methods, etc.) and why these changes will allow us to meet our future challenges. |

|  |  |
| --- | --- |
| **1.3** Platinum Suite | **Understanding the potential - creating a cross-GSS data family: In a huge statistical system how can we solve the problem of finding the right data first time?**  **Darren Barnes (ONS) and Sam Hall (ONS)**  GSS produces a huge range of important statistics that are used by a very diverse audience for very diverse reasons. It is true that users have a hard time locating the statistics they need - even when they know these exist. This is further exasperated by the many platforms the statistics are distributed across. As a community we need to work together to resolve this problem and get our users to the statistics more easily. Sounds easy, right? But how do we actually achieve this? What does achieving this really look like? How do we prepare for the future needs of our users?  These are the questions we need to grapple with. We are keen to share our thoughts about what we can do next and get as much experience and expertise from across the GSS help us develop something really transformational.  *About the speakers:*  Darren Barnes and Sam Hall have worked in ONS for a long time and much of this has been spent in transformational roles. Over the last 12 months they have focussed on what the GSS landscape looks like and how we can improve the situation for both statistical producers and users alike.  **Influences of users of health and care statistics**  **Helen Colvin (ONS)**  The health and care statistics field is complex, with numerous GSS providers producing a huge amount of data and official statistics in different formats, on different websites. This has created confusion with users not knowing what’s available or how best to access it. ONS and our colleagues across government (including Public Health England, NHS-Digital, NHS-England, Department of Health) are driving forwards new ways of working to improve the accessibility of these data sources and statistics to help inform policy, enable analysis and research and improve decision making for health and social care.  Progress so far includes launching GSS theme groups to bring experts together; coordinating the publication of themed outputs of official statistics; harmonizing definitions and improving the coherence of themed analysis; development of a health and care statistics landscape; and launch of a single portal to highlight and document progress. New approaches are being explored by ONS to support this work, including a health monthly publication to provide a summary of health and care data, statistics and analysis produced by the GSS and a health and care statistics compendium providing access to key health and care statistics in one location.  This session discusses the influences of users of health and care statistics in progressing this work and demonstrates some of the digital presentation methods piloted. |

|  |  |
| --- | --- |
| **1.4**  1889 Suite | **Data driven school places, Uganda, & R Shiny**  **Sam Cuthbertson (DFE)**  DfE allocates funding to Local Authorities to build school places where they are needed. I am, however, currently on secondment to an international education charity working with DfID and the World Bank to advise the Government of Uganda on where they need to build new schools. This is being done in a data driven way – hence my involvement.  Uganda was the first country in Africa to deliver on every child receiving a primary education, and it now wants to do the same for secondary. I am leading a multi-organisation team to develop an R Shiny app which visually shows where schools currently are in Uganda, and where there is a need for more school places. As you can tell this work is pioneering and is regarded a leading edge for the sector – I have, for example, presented to the Gates foundation. It also represented a major play for the GSS cross-Government in an international setting to bring completely new insight. |

|  |  |
| --- | --- |
| **1.5** International Bar | **How to build a data science capability and not go mad in the process**  **Tom Ewing (DFT)**  Two short years ago, DfT employed its first data scientist (me!) I was completely new to data science and had no idea what I was letting myself in for. I was also the only data scientist that DfT had and didn’t have anyone to learn from or ask for help. Plus, it turns out that Data Science is actually quite difficult which made the first few months really ‘character building’ to say the least! Over that time we’ve been on quite a journey. We lacked the money, exposure and expertise that a lot of other data science departments had access to, but we did have some things going for us, including an excellent digital / IT department, leadership that gave us freedom to innovate and explore, lots of unrestricted data and access to some fantastic minds in technology, start-ups and transport.  **Data science at MoJ**  **Jon Roberts (MoJ)**  The Ministry of Justice is overhauling our approach to the use, interrogation and visualisation of data as part of a drive to become more ‘data driven’, bringing analytics into the heart of decision making. We are creating a series of interactive dashboards and analytical tools to allow analysts, the public, operational and policy colleagues as well as senior officials to interrogate our data, generate insight and communicate results in a more intuitive and interactive way than traditional approaches.  Within MoJ there is a huge appetite to empower analysts and non-analysts to generate insight from our data and we are building many internal dashboards, covering departmental finance, HR, and operational performance.  Externally, we have published two dashboards (<https://public.tableau.com/profile/moj.analysis#!/>) which focus on Crown and Family courts and have had over 10,000 views in the last 9 months. These allow the user to investigate performance in the justice system, yielding surprising insights into what’s happening in courts around the country and supports our drive to make our published data more accessible than the pdf documents and CSV files we currently publish.  Building on these successes and the team is now running a series of Data Studios, bringing together multidisciplinary teams to tackle some of MoJ’s thorniest problems using Data Science and Advanced Analytics.  We have successfully applied, social media analysis, Natural Language Processing, a range of machine learning techniques and advanced visualisation techniques to generate real business benefit for the MoJ.  Jon will present case studies from the dashboards and Data Studios and describe how the team is driving a completely new way of using data in MoJ, to be more transparent and drive improvements in the Justice System and MoJ more widely. |

|  |  |
| --- | --- |
| **22nd November 3.30pm-4.30pm** | |
| **2.1**  1889 Suite | **Using data science to transform research and innovation policy**  **Juan Mateos-Garcia (NESTA)**  Research and Innovation (R&I) policymakers want to identify and nurture the industries of the future, but these are hard to find in economic statistics that offer a retrospective view of the economy. Big, open and web datasets, data science methods and interactive visualisations and tools which track innovative activity with high resolution and timeliness could address this evidence gap, improving the effectiveness of R&I policy. In this session, I overview Nesta's experience in this area over the last 5 years: the methods, tools and workflows we use, the challenges we face, lessons learned and future opportunities.  **Combining information in multiple spatial datasets to produce estimates and standard errors of population areas**  **Alan Brewer (National Forest Inventory)**  The National Tree Map (NTM©) is a digital map of tree canopy cover covering all land areas in England and Wales generated by Blue Sky International Ltd. using bespoke image-processing techniques of spatial datasets. The National Forest Inventory (NFI) used this product to investigate the extent and character of tree features outside woodland, classified into small woodlands, groups of trees, individual trees and hedgerows. It was found within sample areas that greater spatial accuracy than that achieved by the NTM© map could be attained by direct visual interpretation of aerial photography (hand-mapping) or ultimately by detailed ground survey of the sample area. The NFI accordingly conducted a sample survey of one-by-one kilometre square sample areas outside main woodland in England and Wales within which tree features were mapped by visual interpretation of aerial photography, and a sub-sample was selected for field surveyors to visit to ground survey the sample area as verification and correction of the hand-mapped spatial data.  This paper will describe an analysis technique used to spatially calibrate one dataset with another, and by applying this technique twice; to calibrate the NTM© map to the hand-mapped sample and to calibrate the hand-mapped sample to the ground survey sub-sample, estimates of total areas of classes of tree features can be obtained. The calibration model is of the form:  yj=s\*∑rixij +εj  where the xij are areas of feature classes of the spatial dataset being calibrated within sample square j, the yj are areas of a class or group of features in the calibrating dataset in sample square j, and s and ri are parameters derived from a GIS analysis of spatial coincidences between polygons of the two datasets. Derivation of ground survey-based area estimates and their standard errors through double application of this model will be described and the statistical efficiency of the technique will be discussed. |

|  |  |
| --- | --- |
| **2.2**  Platinum Suite | **Machine learning in production: transforming visa casework and identifying potential victims of trafficking**  **Tom Wakeford (HO)**  The Home Office Data Analytics Competency Centre (DACC) is playing a leading role in the transformation of the Home Office. We have developed a supervised machine learning capability to risk score individuals or entities across different business areas. In this talk, we will describe two case studies. The first is a model to predict the degree of compliance for a particular type of visa cases: a system that will soon be in production. The second is the development of a capability to determine the risk of all incoming travellers to the UK of being a potential victim of trafficking based on historical characteristics of known victims within a tight ‘Window of Opportunity’ between when the pre-departure data is collected and when it is needed for an intervention. To achieve this, we are developing a new state-of-the-art platform to support ingestion and analytics of large volumes of streaming data.  **Exploring mental wellbeing from prisoner case-notes using text mining**  **Jo Lee (MoJ)**  Evidence on people with mental health and substance misuse is costly to commission, yet these individuals are over-represented in the prison population. The most recent Adult Psychiatric Morbidity Survey of prisoners, published almost 20 years ago (Singleton et al., 1998), found that over 90% had one or more of five studied psychiatric disorders (psychosis, neurosis, personality disorder, hazardous drinking, and drug dependence).    To address this gap in knowledge, text mining techniques were explored on prisoner casenotes, where prison officers enter free text to describe prisoners’ progress. The casenotes are written in an ad hoc fashion, recording interactions that range from formal interviews/meetings to chance encounters. The casenotes contain information about prisoner well-being which can be explored to determine mental health despite gaps in timelines for prisoners, a large variation in length or detail of case notes, and no pre-defined coding or structure of information.    Here, I will explore the challenges faced by handling such large datasets (5 million casenotes are recorded each year), and how to explore the relatively vague topic of mental health in text, as recorded by non-specialists. Mental well-being is not easily categorised, but if done properly can provide rich context for prisoners throughout their time in custody. Having encoded mental health issues from the casenotes, we are using them to improve predictive analytics (e.g., the risk of committing violent acts while in custody) and implementing an R Shiny application allowing operational staff to view a summary of issues logged across a prisoner’s case note history. |

|  |  |
| --- | --- |
| **2.3** International Bar | **Integrated Data Enabling Analysis and Statistics (IDEAS): A new data model for integrating data**  **Becky Tinsley (ONS)**  The UK Statistics Authority strategy for UK statistics – 'Better Statistics, Better Decisions’ (October 2014) encourages the bringing together of data to enable better, more timely, evidence-based policy. The Digital Economy Act (May 2017), makes provisions for data sharing and will unlock many of the barriers, enabling a wider range of data to be shared and linked across Government.  So how can the GSS ensure this linkage is done in a consistent, coherent, secure and privacy-preserving way?  The Office for National Statistics (ONS) is developing IDEAS (Integrated Data Enabling Analysis and Statistics) as a change to its data model. Data is referenced to three key spines (about people, addresses and businesses) to enable a consistent, coherent, secure and privacy-preserving way of integrating data.  This presentation will discuss how IDEAS has the potential to provide a consistent basis for delivering statistics and shaping policy across the GSS, and as resource for researchers. It will highlight how we are using what we have already learnt about sources and methods from the work being done to transform our statistical production, and how we want to add value by learning from other integration work already being done across the GSS. We will share some examples of the benefits to the GSS of providing this service. In order to deliver the greatest benefit across the GSS and more widely, ONS will work closely with data suppliers, potential users and other stakeholders in designing and implementing such a model and we are keen to hear your feedback on work so far and planned.  **Household statistics from a future administrative data census**  **Peter Jones (ONS)**  The ONS Administrative Data Census Project is researching an alternative approach to delivering censuses in England and Wales after 2021. Focusing on the use of administrative data and surveys, we have established a series of ‘Research Outputs’ to demonstrate progress towards replicating census topics and producing new outputs after 2021. We invite users to provide us with feedback on our outputs so that we can continue making improvements in response to user needs. We also have three local authority user groups that we work closely with to understand the impact that a potential change in approach would have on the way they use census data.  Due to limited availability of information about households and families in administrative data, we are exploring alternatives to the traditional Census definition of ‘households’. For our Research Outputs we have been using a definition based on the concept of ‘occupied addresses’ and have developed new methods to produce alternative estimates of the number, size and composition of households at local authority level. This presentation will highlight the challenges of using administrative data for these purposes, but also the potential opportunities for combining administrative data with surveys to produce statistics about households that meet user needs. We will also describe the success we have had in warming users up to the potential for a change in definition on this important topic.  Further, the methods and definitions we are exploring have potential to be used more widely across the GSS, increasing harmonisation, frequency and geographic detail of government statistics produced at household level. |
| **2.4**  Directors Boardroom | **Refreshing the Code**  **Penny Babb (ONS)**  The Office for Statistics Regulation has consulted on a new edition of the Code of Practice for Statistics (‘Code 2.0’). We will describe the main feedback to the consultation that ran from 5 July to 5 October 2017. We will share our thoughts about incorporating the consultation comments in the final version of Code 2.0 and provide an opportunity for Q&A.  Statistics are an essential public asset – they provide a window on society, the economy, and on the work and performance of government. They are fundamental to the judgements and decisions made by the public, by government and by an enormous range of other organisations and individual users.  The Code of Practice for Statistics is an important tool in protecting official statistics by setting the standards for producers, to ensure the public can have confidence in them. The UK Statistics Authority has a responsibility under the Statistics and Registration Service Act to prepare and publish a Code and to assess against it. The Office for Statistics Regulation is the regulatory arm of the Authority and provides independent regulation of all official statistics produced in the UK. |

|  |  |
| --- | --- |
| **2.5**  Legends of the Lane | **GSS Good practice team: capability building**  **Martin Ralphs (ONS)**  The objective of the session will be:  a) To raise awareness of how departments are building the infrastructure and facilities that statisticians need to do data science - what is in place, how did they get there, what are the barriers and how were they overcome?  b) To raise awareness of learning and development opportunities in data science - networks and communities, opportunities for training (MSc programmes, ONS Learning Academy / Data Science Campus, Data Science Accelerator, departmental resources) with a strong focus on the practical - what can you do to build your capability, what networks exist that you can plug into?  **Random Forest Analysis: the prediction of fragile A&E systems**  **Ed Beake (DH)**  There are over 100 Accident and Emergency (A&E) Departments in the NHS in England and pressures on A&E Departments are often at their highest in the winter months. There is a belief amongst policy makers that some A&E units are better placed than others to withstand those pressures and we were keen to use our analytical expertise to help identify which units will be most fragile over winter. Conventional approaches involved regression modelling with routine data, or simply assuming that the ones that were considered fragile last winter would continue to be fragile next winter.  Random Forest analysis is an adaptation of the bagged ensemble machine learning technique, based on decision tree models. A decision tree model is termed a ‘weak learner’ in data science. Bagged ensemble machine learning techniques improve the performance of weak learners by constructing many such weak models on the same dataset and combining their results. It is similar to more conventional boot-strapping techniques.  In a recent Department of Health project the Random Forest machine learning technique was used to assist in the prediction of A&E departments that may be considered fragile in the coming winter. We found that the analysis gave new, material insights into the fragility of A&E systems and gave predictions complementary to those given by the models used before.  I will present some of the theory behind and implementation of the Random Forest algorithm using the example of its use in a Department of Health project on the prediction of potentially fragile A&E system. |

|  |  |
| --- | --- |
| **23rd November 10.30am-11.30am** | |
| **3.1**  Legends of the Lane | **Statistical modelling of arrears history data for decision support in Social Housing Management**  **Farida Mustafazade (Newcastle University)**  The use of data to support individual tenant management enables the construction of robust risk assessment tools for Digital Housing Platforms. The question we are focusing on is how to handle tenants in arrears. Landlords need to manage and recover rent debts in the face of changes such as welfare reform. We aim to create prediction models to assist landlords in making decisions on which tenants they need to work with, and how best to work with them.  The exploratory analysis of historical rent balance and arrears history datasets from a few regions across the UK between 2015 and 2017 will be described. These preliminary analyses provide insight into variability in data gathering and arrears management practices. We will go over some of the challenges we faced while using the data from registered social landlords. These include heterogeneity that comes from diverse configurability of the housing management systems and differing reliability. Our overall aim is to construct reliable analytics models able to work equally well with the heterogeneity present in our tenant databases. Once fully developed, this system will enable risk classification to identify tenants most at risk of falling into rental arrears and adverse outcomes to the decision making of income managers in social housing associations. Finally, the ways in which data from credit reference agencies might be used to improve the risk scoring algorithms will be covered.  We gratefully acknowledge our development partner customers for supplying the data.  **Race Disparity Audit: Building a data website and collaborating across Government.**  **Zamila Bunglawala (Cabinet Office)**  The Prime Minister commissioned the Cabinet Office to build an unprecedented data website, with a one-year deadline, to detail Government data on ethnicity. There was no blueprint for the website, therefore the team established two key principles regarding statistical data principles and diverse partner engagement. The Cabinet Office team worked collaboratively with ONS, UKSA, cross-Government data, digital and policy leads, and engaged extensively with diverse external partners from academia, NGOs, local government, open data experts and members of the public, to design, develop and test the website.  The Ethnicity Facts and Figures website went live in October and has been very well received. It is a positive example of data, digital and policy teams working effectively to deliver a key project, and challenges our approach to data analysis, policy development and external/user engagement. This session will share some of the key lessons learned, challenges faced and next steps on the project. |

|  |  |
| --- | --- |
| **3.2**  Directors Boardroom | **Quota sampling guidance**  **Gary Brown (ONS) and Catherine Mottram (DfT)**  A cross-government group of motivated specialists in policy, analysis and methodology have developed guidance on the use of quota sampling for government analysts. This is to provide those commissioning and performing quota samples with an easy to follow pathway to successful survey research. The finished guidance outlines strengths and weaknesses of various alternative sampling methods, what you should consider when deciding on a sampling methodology and top tips for improvement of quota samples. Gary and Catherine will outline the content of the guidance and then explore with attendees whether it would be useful for them in their day jobs.  *About the speakers:*  Gary and Catherine are both members of the small government group who developed the guidance. Gary has worked at the Office for National Statistics since 1999, spending his time working in 9 different branches in Methodology. Prior to ONS, Gary lectured in statistics and ran the statistical consultancy service at Aberdeen University. Catherine is a social researcher and has been a member of GSR since she joined MoJ in 2009. She is now the head of road safety research in DfT. Prior to joining GSR Catherine worked in for a number of social and market research agencies, mainly working for government clients.  **Quality Assurance of Administrative Data (QAAD): The population statistics perspective**  **Philip Humby (ONS)**  Now part of the National Statistics accreditation process, if you produce National Statistics you will need to consider QAAD. The Population Statistics team have been through the process and have been encouraged by the UK Statistics Authority to share their experience across the civil service, so come and find out about how QAAD has been applied to the most widely used statistics in Government.  The Population Statistics Data Administration Team will talk through  • what QAAD is  • how it should be interpreted  • the purpose behind it  • how QAAD fits in with the new Code of Practice  • what QAAD means for you  • the Pop Stats approach  • approaches used elsewhere  • pitfalls you can avoid  • what’s next for Pop Stats and QAAD |

|  |  |
| --- | --- |
| **3.3** International Bar | **The ‘*Data*’ Science of Where - the role of Geospatial analysis in Data Science and Statistics**  **Steve Croney (ESRI)**  Government departments are becoming increasingly data rich, but often remain information poor. Making sense of all this data is a big challenge, but also an opportunity as we witness an increasing convergence of Statistics and Data Science with Geography. This presentation will explore how the Science of Where uses Geography as a common visual language and key enabler to unlock improved understanding of data through; data exploration, analytics, visualisation, sharing & dissemination.  Geographical analysis through GIS can support the production of statistics using spatial statistical tools, as well as help statisticians ‘sell the story’ of the statistics to non-expert groups of users, through insightful engaging visualisations. Big data tools and increased processing power means we are at a significant turning point where it’s far easier to harness the power of location and spatial relationships for generating new understanding, supporting decisions and inform policy.  We will discuss how GIS can help to break down data silos that still exist today within and between Government Departments to make better use of its data, by leveraging this common language of Geography. We will look at examples of where organisations are already taking advantage of Geography in geospatial analysis, visualisation and dissemination.  *About the speaker:*  ESRI build ArcGIS, which connects people with maps, data, and apps through geographic information systems (GIS). Our technology combines maps and data so you can see the world in a smarter way. ESRI UK work with many Central and Local Government organisations helping them to unlock the value in their data through the use of GIS.  **Disaggregating statistics using spatial techniques**  **Chris Gale (ONS)**  At present many ONS statistical outputs are restricted to being published at higher spatial levels, such as at local authority. Providing these data at a finer geographic level has traditionally required the use of small area estimation, based on statistical techniques, to model these data to smaller areas. However, the level of complexity to create such models makes understanding their methodologies difficult. This in turn means anyone without the requisite skills can find it challenging to fully understand the quality of the outputs produced.  An alternative method is using spatial techniques to apportion aggregated data on to smaller geographies. In response to this, ONS Geography initiated a research project to evaluate the potential for a spatial disaggregation technique to be used with ONS produced statistics. The project has been sponsored and seen as a priority by Sarah Henry, the newly appointed Director of Methods, Data and Research at ONS.  This project aims to identify pre-existing spatial disaggregation techniques, analyse the suitability of data sources available to ONS – such as address data, earth observation and big data - and develop a methodology that is a combination of these methods and data sources to create an easily understood spatial method of disaggregating statistics. The outcome of this will be a bespoke methodology and tool-kit that can be applied to ONS statistical releases to disaggregate outputs to a finer spatial level. |
| **3.4**  Platinum Suite | **DCLG’s analytical work in the immediate aftermath of the Grenfell Tower fire**  **Sandra Tudor (DCLG)**  In the aftermath of the Grenfell Tower fire, a first and obvious question was “how many other tower blocks are at risk?” Analysts from across the DCLG group came together quickly and used their various expertise of data sources, data linking and spatial technology to produce robust estimates, which were later verified with directly collected data.  Analysts used DCLG’s English Housing Survey (EHS) to estimate how many other tower blocks like Grenfell Tower there were in England (about 600), where they were (two thirds in London) and who owned them (about 45% by social sector landlords). This analysis was based on rarely used data collected by surveyors during the survey. These findings became the foundation for further work as well as providing much-needed initial estimates for ministers.  Alongside this, we worked closely and effectively with the Homes and Communities Agency to link and use a range of datasets (eg Ordnance Survey, Land Registry, Energy Performance Certificates (EPCs) as well as the EHS). Each additional dataset brought a new aspect to build up the data picture - initially identifying the numbers of high buildings, then drilling down to identify ownership and cladding likelihood. Having delivered this innovative work - linking disparate datasets without common identifiers or standards – the HCA analyst then applied spatial mapping skills and technology to map high-rise locations and their physical characteristics.  In the following days, DCLG conducted a full census of tower blocks in the social rented sector – all local authorities and housing association provided information on their tall buildings and whether they had aluminium type cladding. As the data returns came in, they verified the figures that had been created in the initial analysis by the analysts in those first few days after the fire. The resulting dataset now provides a ‘register’ of at risk buildings (i.e. those with aluminium type cladding) which is driving the testing programme. Processing these volumes of data in a short timescale and with all the inherent issues of a data collection exercise (missing / incomplete values, juxtaposed fields, invalid or incredible values etc) was not easy. One of our data scientists came to the rescue - writing routines to extract and cleanse the data at speed and far more efficiently than using macros.  Such a huge project would normally take months not weeks and is credit to the skills and adaptability of the DCLG and HCA community –who brought together a huge range of analytical and data science skills in a very practical and solutions-focused way. |

|  |  |
| --- | --- |
| **3.5**  1889 Suite | **Data ethics for statistics and research**  **Petros Saravakos (UKSA)**  The statistical system is continuously evolving and over the past years we have developed innovative methods to use already existing data sources, found new sources of data, joined the big data revolution and moved on to even exploit paradata. It is beyond doubt that the statistical system is primed to innovate and embrace data science.  As we continue to explore new methods and utilise a wider variety and a larger volume of data, it is essential that we reflect on the ethical implication of this work. A professional statistical service ought to refine and revaluate ethics in statistics and research. Critical to this will be improving transparency around how we access, use and share public data, in a way that is helpful and instils confidence.  The National Statistician’s Data Ethics Advisory Committee (NSDEC), made up of representatives from academia, the user community, the private-sector and data-owning departments themselves, plays a crucial role in ensuring that - in leveraging the opportunities presented by the data revolution – the statistical community acts responsibly and maintains the trust of its stakeholders and the public.  This presentation will focus on the role and aims of National Statistician’s Data Ethics Advisory Committee and how researchers and business areas can engage and benefit from the Committee’s advice.  **Better Scraping, Better Statistics? Using web-scraped data in statistical outputs**  **Matthew Greenaway (ONS)**  Web-scraping – the collection of data automatically from the internet - is a data science technique which provides an exciting new data source for government statistics. Harnessed correctly, web-scraped data may help us measure important elements of the modern economy in a more timely, granular or complete fashion, something which is a priority for the Office for National Statistics (ONS) following the Bean review of Economic Statistics. However, web-scraping can be fraught with challenges and, if done badly, can expose organisations to legal or reputational risks. Using web-scraped data to produce statistics also presents numerous methodological challenges such as addressing coverage and accuracy.  This presentation will summarise how ONS have responded to these challenges. The first half of the presentation will cover how we have worked with the National Statistician’s Data Ethics Advisory Committee and legal experts to develop web-scraping guidance which will ensure our web-scraping is carried out consistently, ethically, and cognisant of relevant legal issues. The second half of the presentation will detail several examples of how we are implementing web-scraping in the domain of economic statistics – such as, for example, improving our job vacancy statistics by using data from jobs portals - focusing in particular on the relevant statistical challenges.  We will conclude by summarising our experience and offering advice to other GSS statisticians who are interested in using web-scraping to improve their statistics. |

|  |  |
| --- | --- |
| **23rd November 11.45am-12.45pm** | |
| **4.1** International Bar | **Improving data communication, 90+ tables to an API**  **Hiren Bhimjiyani (BEIS)**  Quick and easy access to data is an important aspect for government statistical publications. How people use the data we publish can never be fully anticipated, so we should make it easy for them to take our data and reshape it for their needs.  Traditionally, published data has been accessible in tabular format via spreadsheets or csv files. Unfortunately, with this type of format, accessing and using the data to its maximum potential can be limiting and difficult to accomplish without significant efforts in data wrangling, particularly when trying to access time series information on individual variables across many spreadsheets.  This presentation is about how the Data Science Team in BEIS has improved data communication by transforming 19 years’ worth of published data from the Digest of United Kingdom Energy Statistics (DUKES) held in spreadsheets across 90 tables, and converting it into a structured linked dataset which can be accessed publically through the BEIS (CBAS) API. <http://njs.analysisoncbas.co.uk/energy/data>  Publishing data in this form has many advantages such as allowing users to select the cuts of data they need, formatted in a way that suits them, and enabling them to get on with the analysis they need to do. In addition, publishing the data like this opens it up to a wider community of developers, expert users, as well as enabling interactive data visualisation and app development.  Without access to the raw data, the work required to transform published data into a structured database, suitable for delivery via an API, can be a challenging task for the team publishing it. We will share our experiences of working with such a team, reshaping the data (and what this shape should be for maximum flexibility), building an API for delivery, and how to begin to identify the user needs. We focus on the lessons learnt and how we delivered a usable product, rather than technical details or abstract architectural theories, in order to show how we met user needs at all points of the project.  **The use of Data Science to improve the way we process and disseminate data**  **Tom Davies and Mark Burley (DWP)**  Purpose: The purpose of this work was to use data science techniques to improve the accessibility, dissemination and usability of our statistics.  Method: Utilising a combination of HTML, CSS and JavaScript programming, we have generated a series of interactive dashboard-style visualisations for our Official Statistics. These provide another format to how we release our statistics to accompany existing formats such as our First Releases, ODS tables and Stat-Xplore – our tool that allows users to create bespoke tables. The interactive visualisations summarise vast amounts of data into an understandable, user-friendly format. Built using API’s, the dashboards collect and process detailed data much quicker than ever before and empowers the user so they can retrieve the data they want. With accessibility in mind, we ensure that our dashboards are accessible through computer, tablet or mobile.  Results: We now have a suite of dashboards covering 6 key Departmental statistics, publically available over the internet – here’s a link to examples (NINO stats and Fraud stats). This has made statistics much more accessible to users than previously – the information in the interactive map would otherwise have to be presented in a spreadsheet of more than 1,000 rows – which lots of users wouldn’t find accessible. We’ve had feedback from various users that these interactive visualisations are really beneficial. For example one of our DWP Operational manager users said; “I can now drill down from high-level strategic data to low-level local data in 3 clicks of a mouse, and the information is absolutely clear so I know I’ve got the right answer”  By sourcing these visualisations from APIs, this has freed up resource from manually updating, to use on exploring the application of other Data Science techniques such as automating statistical publications. |

|  |  |
| --- | --- |
| **4.2**  Platinum Suite | **The Dissemination Game: How to communicate official statistics to non-expert users**  **Hannah Thomas (Gov Wales)**  This presentation will look at the approaches that can be used to improve the communication of official statistics to non-expert users.  I will draw on my four years of experience at the UK’s Office for National Statistics (ONS), in particular the three years I spent in the Digital Publishing Division creating content for a website called visual.ons.gov.uk (a sister website to the official ONS website).  The material in this presentation is subjective. How to win at the dissemination game is an ongoing challenge and people across the globe have different opinions on what works and what doesn’t. Therefore my aim is not to preach, it is to share my experiences as a statistician working in the digital area of an official statistics office.  **Understanding crime in prisons: the problem**  **Jack Tattersall (MoJ)**  The number of crimes committed in prisons has increased dramatically in the last few years, a fact that has not been lost on the media, the public, and the government.  Whenever such figures are published we are treated to what has now become a (not unjustified) yearly national outcry about the “UK prisons crisis”. Tackling this crisis was one of the key pledges laid out in the 2016 Prison Reform white paper.  It was against this backdrop that we were commissioned by the Prisons Security Board to investigate how we can use statistics and data science to understand more about this crisis, and ultimately to contribute to solving it.  What we’re doing  Working closely with colleagues in the Prison Reform Policy team, I’ve been developing an R Shiny app to help visualise and understand what’s happening in our prisons.  Some of the things that the tool can be used for include:  • Identifying priority prisons that are most in need of reform;  • Seeing what crimes are being committed in those prisons;  • Displaying demographic information about who is committing them;  • Examining how these cases are handled by the justice system.  This will enable our customers to get a clearer picture of what’s going on in prisons, will enable them to make better use of our data, and will make it easier to form policy on the basis of solid evidence. |

|  |  |
| --- | --- |
| **4.3**  Legends of the Lane | **Cracking address matching: Why it matters & why it’s not easy**  **Alistair Calder (ONS)**  As in 2011 the address register is right at the centre of the 2021 Census design and impacts on virtually every census process. A high quality address list is used to deliver internet access codes, to allow targeting of follow-up for non-responding households and to underpin estimation and the production of outputs.  Our new approach will feed in address-level intelligence from administrative, commercial and open-data sources enabling much better targeting of resources. The address register also forms one of the key spines for the integration and reuse of administrative data – part of a complete redesign of how ONS manages, processes and analyses data.  High quality matching is essential throughout. ONS have already developed expertise in matching between addresses and, following discussions with Government Digital Services (GDS), have been leading on developing an address look-up and matching service for use across ONS – and potentially across government.  A combination of traditional matching and data science techniques are being applied. A parsing algorithm splits the address entered into its constituent parts (eg. building number, street name and town name) while machine learning has been used to build lightning fast Elastic Search indexes.  This session will describe current thinking for the census and beyond and describe the innovative work currently being carried out applying data science thinking to address matching.  **Use of the VAT data within ONS, in collaboration with the ONS Data Science Campus**  **Andrew Sutton (ONS)**  VAT returns provide a timely feed of information on the turnover and expenditure from a massive sample of individual companies. Clearly this data set contains a wealth of economic information, as reflected by its use within ONS economic statistics. VAT based estimates of short term output indicators are being developed, and these will soon feed into estimates of quarterly GDP.  Here we present a different use of the VAT data within ONS. In collaboration with the ONS Data Science Campus, we have been investigating the potential of VAT returns as an early indicator of the direction of changes in GDP. Starting from a simple diffusion index, we have constructed an experimental, reactive, VAT-based early indicator of GDP. In this talk we discuss issues around the VAT data source and the building of the index, before presenting results coming out of this pioneering work. |

|  |  |
| --- | --- |
| **4.4**  1889 Suite | **Global GSS: International priorities, preparing for Brexit, and International Development**  **Emily Poskett (ONS), Robert Bumpstead (UKSA), Will Laffan (UKSA) and Ian Cope (ONS)**  The vote to leave the European Union has raised questions about how the UK should prioritise international engagement in a post-Brexit world, and how the GSS can make an impact internationally. Chaired by Robert Bumpstead (UKSA), this session will cover international issues and priorities. The GSS International Committee (GSS IC) aims to explore the international issues of concern to each department, as well as the new international statistical relationships that will be important post-Brexit.  In this presentation, the Chair of the GSS IC, Ian Cope (ONS) will provide an overview of the role of the committee and its aim to support the GSS to engage globally. This will be followed by a presentation by Will Laffan (UKSA) on preparing for Brexit and the implications for the UK statistical system. The final presentation will be delivered by Emily Poskett (ONS), who will discuss the aims and achievements of the International Development team, the new DFID/ONS collaboration and their work in partnering with Rwanda, Ghana, Kenya and UNECA to drive statistical modernisation, as well as opportunities to get involved. The session will end with a Q&A session. |
| **4.5**  Directors Boardroom | **Discover how design of local policy is elevated by engaging data analysis: a LEP perspective**  **Lisa Clark (Sheffield City Region Executive Team)**  Producing statistics for wide aspects of economic policy is highly important, but making such evidence insightful and compelling is crucial. It enables stakeholders to visualise the critical issues and allows for easier decision making. Furthermore, it facilitates policy design for clear implementation and direct purpose. In this presentation, Lisa will share her experience as to how Local Enterprise Partnerships can transform dull data tables into engaging statistics and graphics. A breadth of examples will be presented, so the audience can appreciate the data journey and gain insight into what does (and doesn’t) work.    *About the speaker:*  Lisa is Senior Economic Policy Officer at Sheffield City Region. She is an applied data scientist, with extensive experience of mathematical modelling, computer simulation and data mining. She has specific application in the areas of economic policy and devolution.  **Why should we have a ‘place-based’ industrial strategy anyway?**  **Paul Swinney (Centre for Cities)**  Different parts of Britain offer very different advantages as a place to do business. And this influences the clustering of the national economy in very specific places. This presentation will explore the geography of jobs and businesses, why they cluster and what this means for the industrial strategy. |

**Exhibition Stalls**

**ESRI(UK)**

Esri is the global market leader in geographic information systems (GIS), offering the most powerful mapping and spatial analytics technology available. Since 1969, Esri has helped customers unlock the full potential of data to improve operational and business results.  Today, Esri software is deployed in more than 350,000 organisations including the world’s largest cities, national governments, businesses, and academia.  Visit our exhibition stall and attend our presentation to hear about the power of Geography in Statistics and Data Science.

**Royal Statistical Society**

The Royal Statistical Society is one of the world’s leading professional bodies for statisticians and data analysts. We have a strong relationship with the UK Civil Service, providing government statisticians with a collective voice to help shape policy, a forum to network with government and non-government statisticians nationally and locally, and help with professional development. Visit our stand to find out how to join our community, get involved in our work and apply for one of our professional awards.

**Data Collection Transformation Programme**

The Data Collection Transformation Programme seeks to rebalance Office for National Statistic’s data collection activity significantly toward wider, more integrated use of administrative and other non-survey data sources, thereby reducing our reliance on large population and business surveys. We are in the process of doing this, by:

1. moving survey data collection online, changing existing processes so that survey data is predominantly collected using online methods rather than existing use of paper, telephone and face-to-face interviews; non-online methods would only be used where there is an exceptional reason to do so;
2. survey and administrative data integration will allow us to make greater use of existing non-survey sources and reduce the size of remaining survey samples and the number of variables that require a survey response;
3. moving to generic, shared IT systems; integrating our systems so that we have common platforms, for example, having a single online collection tool;
4. business and social survey rationalisation, bringing surveys together so that we only ask for a piece of data once, regardless of context;
5. field force modernisation of contracts and procurement of census services.

If you would like to know more please come see us at the event. Jack Sim and Vicky Palmer will be present and able to assist.

For additional information before or after the event, please contact the programme office: [DCTP@ons.gov.uk](mailto:DCTP@ons.gov.uk)

**Administrative Data Research Centre for England**

The Administrative Data Research Centre for England (ADRC-E) is part of the ESRC-funded Administrative Data Research Network, a UK-wide partnership between organisations with world-class expertise in using and analysing administrative data. The Centre carries out administrative data research and provides facilities for accredited researchers to access linked de-identified data from government and non-government administrative sources. The ADRC-E also offers a range of professional short training courses to support researchers working or wishing to work with administrative data.

**Committee members**

Chair: Alex Miller, Department for Education

Sophie Stewart, Department for Education

Alix Crabtree, HM Revenue and Customs

Juwaria Rahman, Her Majesty’s Inspectorate of Constabulary and Fire & Rescue Services

Sarah Murphy, Ministry of Defence

Richmond Davies, Information Services Division of NHS Scotland

Katy Morgan, HM Revenue and Customs

Ralph McDevitt, Department for International Development

Clare Betts, Department for Environment, Food and Rural Affairs

Sarah Tucker, Office for National Statistics

Sean Mattson, Department for Business, Energy and Industrial Strategy

Michael Cole, Department for Work and Pensions

Ian Hillis, Insolvency Service

Duncan Cook, Department for Communities and Local Government

Pennie Vargas, HM Revenue and Customs

Chris Stickney, Office for National Statistics

Stefanie Taylor, Qualifications Wales

Alison Eve, UK Statistics Authority

**Conference blog and twitter feed**

You can find our official blog for the GSS Conference at <https://gss.civilservice.gov.uk/blog/category/gss-conference-2017-blog/> . We have been posting in the run up to the conference, and will be posting during and after the conference too!

For tweets about the conference, see the official GSS Twitter account (@UKGSS). If you send any tweets about the conference, please use the hashtag #GSSConf.

The [GSS Slack network](https://gov-stats-service.slack.com/) has a dedicated [#GSSConference](https://gov-stats-service.slack.com/messages/C6LRU8N01/details/) channel so you can begin to network with colleagues across departments before the day. The conference team will also be using Slack to post communications throughout the 2 days of the conference to ensure that everything runs as smoothly as possible.

To join the conversation:

• Go to the [GSS Slack log in page](https://gov-stats-service.slack.com/).

• Register for an account using your Government email address.

• Once logged in open the list of “Channels” (using the “+” icon) on the left hand side.

• Choose #GSSConference and you’re in!

If you experience any problems the [Good Practice Team](mailto:goodpracticeteam@statistics.gov.uk) are available to help you. We appreciate some department’s IT restrictions will mean being unable to join from a work PC, but hopefully you can still get involved using a mobile device.