DATA
FOR
POLICY
2015

Policy-making in the Big Data Era: Opportunities and Challenges

15- 17 June 2015
University of Cambridge

Book of Abstracts

DATAFORPOLICY.UK @DATAFORPOLICY #BIGDATA4POLICY

List of Presentations

(In the same order as they appear in the Conference Programme)

Day 1 – Monday, June 15

Ten challenges of Big Data for Social Science

Kenneth **Benoit**, Professor of Political Science Research Methodology and Head of the Department of Methodology, London School of Economics and Political Science (LSE)

The ubiquity of "big data" about social, political and economic phenomena has the potential to transform the way we approach social science. In this talk, Professor Benoit outlines the challenges and opportunities to social sciences caused by the rise of big data, with applications and examples. He discusses the rise of the field of data science, and whether this is a threat or a blessing for the traditional social scientific model and its ability to help us better understand society.

Do Bots of a Feather Flock Together? Analysing the role of Twitter bot behaviour in political discussion

Alison **Powell**, LSE; Nick **Anstead**, LSE; Les **Carr** University of Southampton; Susan **Halford**, University of Southampton; Dhiraj **Murthy**,Goldsmiths College; and Mark **Weal**, University of Southampton

Recent estimates suggest that approximately 8.5 per cent of active Twitter users are actually bots, which create 24 per cent of the tweets on the social media site.
This large number of bots become especially problematic when social media data is used to aid in policy-making or make claims about public opinion. In these arenas, it is assumed that information sharing on social media is the product of human action. The prevalence of bots, however, undermines this assumption. Given that public discourse, media representation and to an extent public policies increasingly draw on data from social media, it is important to understand the role of bots in this information ecosystem.

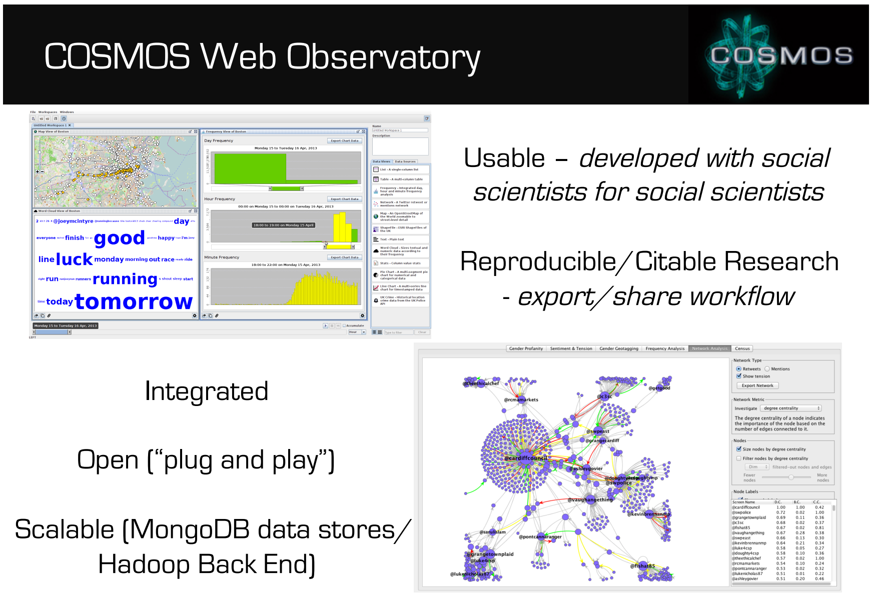
This paper investigates how bots operate as information sharing entities, describing their role in the information sharing ecosystem and their political economy. To do this, we employ an experimental release of bots into an existing discursive community focused on a recurring hashtag. Bots of different service levels and functionality are sourced and released to follow two Twitter accounts tweeting using the hashtag linked to the weekly Prime Minister's Questions (#pmqs) for 4 weeks. The patterns of linking and information sharing behaviour linked with the hashtag for these (experimental) weeks are compared with baseline data gathered for the previous 4 weeks by one of the research team, using Flow140 and network analysis tools. The results highlight the extent to which social media information ecosystems are increasingly hybrid spaces where the behaviour of non human agents can be as significant as human actors.

In both the case of Prime Minister's Question Time and more generally, this research raises interesting questions for political reporting in a hybrid information environment and for practices of public policy making, when decision makers are seeking to harness social media data to make more informed choices.

Computational Cyber and Human Security Analytics using Big Data

Peter **Burnap** and Matthew **Williams**, Cardiff University, UK

There is an established interest in applying digital methodologies to support social media analysis for the understanding of community networks and cohesion, communication patterns and topic-specific sentiment and opinion, for the benefit of society, as well as for financial gain. We present the Collaborative Online Social Media Observatory (COSMOS), a distributed digital social research platform, providing on-demand analytics for the purposes of observing and inductively interpreting socially significant evidence gathered via the emerging uptake of social computing (e.g. Twitter, Facebook, Google, Blogs, News reporting agencies). The COSMOS software platform reduces the technical and methodological barriers to accessing and analysing social media and other forms of open digital data and is set apart from all other existing social media analysis software due to its novelty in five areas: i) it is supported and informed by rigorous methodological and technical research conducted by an interdisciplinary team (computer and social scientists) that informs users in their analysis; ii) the platform allows for the linking of multiple digital data sources (social media, other digital, curated and administrative); iii) it integrates a number of data analysis tools using a workflow model (e.g. sentiment analysis can be followed by a social network analysis, which can then be geo-located); iv) its analysis algorithms are open, transparent, inspectable and refreshable/adaptable by users; and v) users do not need any knowledge of programming.



COSMOS Web Observatory

Usable – *developed with social scientists for social scientists*

Reproducible/Citable Research - *export/share workflow*

Integrated

Open ("plug and play")

Scalable (MongoDB data stores/ Hadoop Back End)

Burnap, P. et al. (2014) '*COSMOS: Towards an Integrated and Scalable Service for Analyzing Social Media on Demand*', International Journal of Parallel, Emergent and Distributed Systems

Figure 1 – COSMOS Platform Snapshots

The volume of data produced on a daily basis requires significant computational resources to analyse. For example, COSMOS collects 3.5–4 million tweets a day via the Twitter Streaming API. The collection has been operating around 3 years and has collected more than 2.5 billion tweets. It also collects every geo-coded tweets published within the UK and has an archive of around 250 million geo-coded tweets. COSMOS also allows users to collect bespoke datasets from Twitter, using topic specific keywords, to collect data surrounding real-world events, which has resulted in an archive of tweets particularly focussed on events that could impact human and social security and safety – for example, the reaction to the Woolwich and Boston terrorist attacks in 2013, the victory of Obama in 2012, targeted hate speech toward high-profiles individuals, and celebrity suicides. COSMOS supports a number of forms of analysis in an integrated environment such that the output of one form can feed into another in a workflow. Analysis methods include the classification of: gender, age, location occupation and social class, sentiment, tension, topic detection and social network analysis and hate speech (see Figure 1). The first part of the presentation will introduce the COSMOS platform, while the second part will focus on our substantive research in the area of computational cyber and human security analytics.

Open and widely accessible social micro-blogging technologies, such as Twitter, are increasingly being used by citizens on a global scale to publish content in reaction to real-world events. The diffusion of this information following events can manifest itself in a number of ways, ranging from supporting social resilience through calls for assistance or help, and spreading of advice and information to the socially disruptive, by providing a platform for the distribution of misinformation, rumour and antagonist commentary. When a Twitter user reads a tweet, they can perform an action to 'forward' that tweet to all other Twitter users who *follow* them, incrementally widening the potential readership of the original tweet. This information can contain textual content, hashtags and URLs, from which a variety of temporal, content and social metrics can be derived for modeling purposes. This modeling is important to governance in a digital era, to understand how long a piece of antagonistic text might continue to be propagated in the *Twittersphere* as it may pose a risk to social cohesion. Likewise, it could be important to understand the factors that are likely to lead to information from official governance sources, such as law enforcement, reaching a large number of people in a short space of time. One COSMOS case study focused on information flow propagation in the aftermath of a terrorist event, with an objective of producing statistical models that could support the inclusion of 'new media' as a source of evidence when developing policy on how to observe online events and when/if it is ethical and necessary to intervene with, for example, a counter narrative. Using the propagation of 'cyberhate' as an example, we discuss the issues with representing results (statistically and visually) from such a study and also the barriers to intervention when considering freedom of expression.

The study can be considered as an aspect of the broader study of information diffusion. Understanding the diffusion of information in social networks has received significant attention from a topological and social influence perspective, but as a recent survey identified, further research is required to better understand the significance of opinion and social factors. Therefore, our key research question is – what are the opinion (sentiment) and social factors that predict large (the size) and long lasting (the survival) information flows following a terrorist event? We present the results of this case study (for example, see cumulative survival of tweets based on agent type over time in Figure 2). Further to the information flow analysis we also present insights into online interactions and behaviors for the purposes of understanding how society reacts to real-world events that could lead to problems such as social disorder, wellbeing concerns, or further disruptive events. Using natural language processing to extract predictive features from Twitter data, we will demonstrate: a rule-based approach to detecting social tension indicators over time following a high-profile case of alleged racism in sport; a linguistic analysis of cyber hate speech; and an ensemble machine learning classifier to identify and categorize suicidal communication – including modeling contagion of suicidal ideation through Twitter.

Big Data, Twitter and the Scottish Referendum

Thomas Hunter **Smith**, Office for National Statistics, UK

This report looks at mining of social media data to follow demographic trends - in particular attempting generate a comprehensive social commentary in the build up to the Scottish referendum. Up to date Twitter data on the referendum was extracted and analysed through the use of adaptable and potentially powerful data mining tools which once refined, could in future be used to provide insight across a variety of settings. We will explore how the generation of Word Clouds, dissemination of Geo-Location information and application several methods of Sentiment Analysis could be used channel vast amounts of open data into a robust and timely method for modelling demographic change.

Social reflection of new policies in the big data era

Jonathan **Bright**, Helen **Margetts**, Scott **Hale** and Taha **Yasseri**, Oxford Internet Institute, University of Oxford

In this paper we explore the ways in which data generated by social media platforms can be used to understand the social impact of policy change. We use as a case study the UK Department of Work Pensions, where far-reaching changes to the UK benefits system are currently underway, specifically the introduction of Universal Credit and Personal Independence Payments.

We argue that social media data can be useful for policy analysis in two key respects. First, these data can provide ways of measuring public awareness of and attention to policy changes, as well as showing the information sources that citizens are using to find out above them. Second, they can provide indications of public reactions to specific policies, and public experience of services.

Methodological development in the use of social media data for policy analysis remains in its early stages. However, we highlight potential data sources and analytical methods such as:

- Google Trends data provides useful indicators of how many people are thinking about a topic at any given time, and the extent to which the public is aware of new policies (for example, Universal Credit), and may also give short

term indications of upcoming changes in, for example, the number of jobseekers.
- Google Search data provides useful indicators of where the public gets information on particular DWP policies. This could allow the DWP to fine tune its communication strategy, and also tell them which other websites are currently informing the public on DWP policies (including, for example, private sector sites over which DWP has no control).
- As well as Google, Wikipedia provides a rich source of information on how many people are interested in a given policy or initiative at a time, or the extent to which public information is available and shared via peer production.
- General social media platforms such as Twitter and Facebook provide a means of assessing how many people are discussing any policy at a given time, and also a way of potentially evaluating their sentiment towards that policy. These platforms can provide a useful indication of the impact of specific media events or press releases.
- The DWP's own social media accounts, run by their network of local Jobcentres, provide a further resource in this regard. They may be useful both for recording feedback and contacting specific regional populations. However, the overall level of activity around these accounts varies (particularly in the early days of policy change), meaning this method should be used cautiously.
- We also highlight the importance of "themed" social network sites such as Mumsnet, though these sites did not fall in the scope of our empirical analysis.

Some of the above can also be analysed to supplement or ultimately replace more traditional – and far more costly - techniques such as survey analysis. In comparison with surveys, we argue, social media data present the following advantages:

- They are comparatively cheap to collect, when compared with the cost of traditional sample surveying.
- They can be collected and analysed quickly, once systems are put in place which perform such analysis.
- They offer potentially very large samples, which means that questions about smaller "subgroups" can be accessed (for example issues affecting a certain geographical area).

However we also argue that caution is needed in interpreting results derived from social media data. Social media sources do not immediately replace other sources of social research data such as sample surveys: the science behind these methods is still developing; and major questions remain around how to employ them properly. In particular, the following challenges are important:

- Social media users are not representative of the population at large (their use is much more widespread amongst the young, for instance).
- Different people use these media in different ways: most of the postings to social media are made by a subset of overall users, who again are unlikely to be representative.
- Techniques for the automatic extraction of opinions from social media and analysis of sentiment are still developing and have to be interpreted with caution.

Overall, we argue that as social media become more deeply embedded in the fabric of life, it is increasingly difficult to ignore the potential they present to social research which can inform policy-making and service delivery, providing data in both quantities and richness that would be prohibitively expensive to duplicate with traditional survey research. However, further research of this kind is needed before social media data can be integrated into working practices. In this context in particular, we recommend that all social media data be “benchmarked”, as most indicators developed in the report are difficult to interpret in isolation. This is a process which may involve:

- Comparing social media indicators over time (for example, if positive ornegative sentiment is increasing or decreasing).
- Comparing social media indicators across Departments or policies (for example, seeing if one Jobcentre is attracting more social media attention than another).
- Linking these data with other sources of socially generated information, for example web traffic to DWP’s own information and jobs websites.
- Linking these data with other trusted sources of information generated with more traditional techniques (such as survey data and ONS employment statistics).

Big Data in Telecommunications

Mustafa **Ergen**, Turk Telekom Argela

Mobile and fixed operators are building big data platforms to analyze the traffic traces of subscribers. By analyzing the packet traces they can predict the user-level experience of the application. Based on this analysis operator can understand that for example the user-experience is poor do to high round-trip delay for the video clip. The operator can then cache the video-clip and hence improve the user’s experience.

By quantifying the location where the main services that users are interested in, operator can make recommendations of which locations to add capacity and/or place wireless hotspots, which locations are having issues even with mobile penetration, eg. due to bad setting of radio parameters. By geo-locating applications and looking at temporal trends (daily, weekly) operator can provide insights into cultural issues which can result in more targeted marketing campaigns in different regions.

Big Data Empowered Self Organization For Split Plane 5G Cellular Networks

Syed Ali Raza **Zaidi**, University of Leeds; Mounir **Ghogho**, University of Leeds; Muhammad Ali **Imran**, University of Surrey; and Selcuk **Bassoy**, University of Surrey

The so called “scissor-effect” is triggered by the gap between the costs for providing wireless connectivity and the generated revenue. This has motivated both operators and vendors to adapt innovative strategies for their 5G network roll outs. Innovation is required not only in terms of the network architecture (to deal with the “exabyte flood”) but also in terms of business models. It is envisioned that significant additional revenue can be generated by introducing new value added services (VAS) in conjunction with intelligent business models. From the operator’s perspective, these services can be introduced with zero or minimal increase in capital and operational expenditures. In other words, these VAS seek to maximize network and infrastructure utilization. This has led to a paradigm shift in terms of the network architecture for the 5G wireless network. More specifically, a fluid/elastic network architecture which supports “user centric” services will be a central feature for the 5G wireless networks. Such a fluid architecture leverages the fact that for ultra-dense networks both the control and the user-plane can be decoupled. Consequently, network can be treated as a soft entity which can support variety of user demands by dynamic reconfiguration of the user plane. Furthermore, activities across heterogeneous control planes can be coordinated through the network state awareness and resource virtualization. In brief, 5G networks are envisioned to take a soft and reconfigurable design approach for decoupling the infrastructure utilization from the network/resource utilization. It is difficult if not impossible to realize such an architecture without the real time network and the user state data. With around 6.2 billion internet connected devices by the end of 2020, the generated data has enormous volume, high velocity and huge variety. In other words, the generated data can be easily considered as one of the most important examples of “big data”. In this article, our main objective is to highlight how networks can be self-organized by leveraging big data empowered self-organization specifically for the split-plane architecture. To the best of authors’ knowledge the design framework for the big data empowered self-organized networks under split-plane architecture remains unexplored to date.

Planning for sustainable urban transport: An activity-based model with large-scale travel diary, POI and review sites data

Lun **Liu** and Elisabete **Silva**, University of Cambridge

Research Questions:

Transportation accounts for ca. one third of all energy consumption on earth, of which sixty percent is generated from people’s daily travel. As a result, tons of research has been dedicated to modeling the relationship between the built environment and people’s travel behavior with the aim of making transport more sustainable. The field of transport modeling has witnessed a slow but constant shift from conventional “four-step” travel demand models to the “activity-based” approach in the last thirty years. The advantages of activity-based model lie in a deeper understanding of the motivation of travel, as well as a multi-faceted and more detailed description of travel behavior such as trip chaining in one tour. Despite of the modelling sophistication, it is said that some countries such as the US has reached the point where the majority of new transport models projected are activity-based.

However, most existing models consider only the one-way influence of the built environment on the travel behavior, while in fact the travel behavior could also exert a feedback on the built environment through the location and relocation of facilities in the long term. Though such a feedback loop has been proposed before (Timmermans, 2010), it is seldom simulated at the city scale. Part of the reason lies in the lack of relevant data. The availability of Point of Interest (POI) data now enables such an analysis by providing a yearly or even monthly update of the spatial distribution of various facilities, with which the feedback effect can be examined. Furthermore, the residential and travel behavior differentiation of various social groups leads to the spatial differentiation of high-end to low-end facilities. This effect can also be analyzed by identifying the price range of facilities with the data crawled from review sites. Thus, the big data functions as a key supplement to traditional travel diary data to provide a new approach in modelling this bi-directional interaction. The research aims to answer:

1. How does people's activity-travel behavior interact with the built environment, especially the location of various facilities?
2. How does this interaction differentiate among various social groups?
3. How can land use planning help reduce car use, thus promoting sustainable urban transport?

Data:

The research uses Beijing as the case of study, which is a fast-growing city confronted with various transport problems. Three data sets are applied. (1) A city-wide 24-hour travel dairy survey in 2011 on ca. 30,000 households, containing ca. 250,000 trip records. (2) City-wide POI data in 2011 and 2014 on almost all types of facilities including office buildings, supermarkets, stores, bars & restaurants, entertainments, hotels, banks, hospitals, schools, parking lots, etc., containing ca. 100,000 points. (3) Average expenditure per capita in commercial facilities crawled from Dazhongdianping, the Chinese version of Four Square.

Methodology:

The research develops an integrated activity-travel land use model (Beijing Activity-based Transport-Land Use Model, BATLUM), which can be applied to test the impacts of various policies on urban transport. Particularly, the model deals only with nonwork travel which accounts for seventy percent of all trips, considering the quite different nature of work and nonwork trips. BATLUM is composed of three sub-models. (1) The Household Generator produces a synthetic population with spatial distribution for each simulation year. It is innovative in taking into consideration the fact of residential differentiation among social groups, which is further linked to the location and relocation of various facilities. (2) The Person Day Activity-Travel Planner takes the configuration of DaySim (Bradley, Bowman & Griesenbeck, 2009), which simulates people's decision making on primary activity destination, tour stops, travel time, mode choice, etc. based on utility maximization. Comparing with existing models, this model takes into account the intra-household interaction in activity-travel and a more detailed classification of activity types, enabled by the detailed types of facilities in the POI data. Moreover, instead of applying the same decision rule on the entire population, the model allows for varying decision mechanisms among social groups, considering their respective spatio-temporal constraints in activity-travel, which is further linked to the distribution of high-end to low-end facilities. (3) The Facility Location Choice module decides whether new businesses move in and whether existing businesses quit at each location based on the number of customers, the concentration of competitors, as well as planning restrictions. This in turn affects people's activity-travel until equilibrium is established. The model can be applied to examine the impact of various planning and transport policies, including the land use mix, density restrictions, affordable housing developments, transport infrastructure provision, taxes and incentives, etc. The calibration is conducted from a base year of 2011 to 2014, with the travel dairy data in 2011 and the POI and review site data in both years.

Key findings:

The research is still on-going and we expect three key findings. The first is a better understanding and quantification of the feedback effect of people's travel behavior on the location of facilities. If it turns out to be significant, it would highlight the importance of accommodating this effect in transport modelling. The second is an understanding of how different social groups react to the built environment differently. The third is a policy-support model for evaluating the impacts of various planning and transport policies, which is a refinement of existing tools by a more realistic and comprehensive simulation of activity-travel with big data.

The Cabinet Office, ethics and data science

Paul **Maltby**, Director of Open Data and Government Innovation at the Cabinet Office
Cat **Drew**, Senior Policy Adviser, Data Science at the Cabinet Office
Simon **Burall**, Head of Dialogue at Sciencewise Expert Resource Centre

Sciencewise and the Cabinet Office Data Science Team jointly propose a panel discussion on the Data Science Team's work to develop an ethical framework for the use of data science in government. Sciencewise is supporting the Cabinet Office's work on the framework, which is aimed at both policy makers and analysts in government.

The government is looking to expand their use of data science and encourage policy makers to make use of the opportunities provided by the new technology. Data science can aid government's work in terms of both operations and research, by helping to make services more efficient, generate insights and save government money.

However, government is aware that this is new territory, throwing up ethical questions. As data science is emerging as central to our future economy, pioneered by big tech companies like Google and Facebook, government has a responsibility to take a lead on the ethics of data science, and clarify the need for ethics that go beyond the law. Many of the questions facing the Data Science Team at the Cabinet Office will be familiar to conference attendees: they relate to trust, transparency, fairness, accountability, privacy and consent.

The framework, currently under development, outlines the benefits and the limits of data science, clarifies the need for ethical guidelines beyond the law, urges policy makers to consider ethics throughout the project, not just at the end, and consider public opinion. The framework also encourages policy makers to innovate and make greater use of data science.

The Cabinet Office Data Science Team has consulted with experts external to government to better understand established thinking on ethics, the context of developments in data science across the public and private sectors, and the barriers to greater use of data science by government. The views of experts have been sought from the worlds of academia and civil society, including privacy groups.

As part of their wider work on data, the Cabinet Office has also been undertaking an open policymaking process on aspects of data use in public services. This process has seen the Cabinet Office convening a space for civil society engagement with government departments to explore the benefits, risks, limitations and governance for better using data, including personal data, within government. This process has been conducted in a transparent way with key meetings and documents covered on the website http://datasharing.org.uk/ and is something that we would like to look to see how we might learn from for data science.

Ethics in the Age of Big Data

Ross **Anderson**, Professor of Security Engineering, Computer Laboratory, University of Cambridge

It takes about fifteen years for lawmakers to catch up with technology, so a prudent engineer will not just ask whether a new project is legal now, but whether it will still be legal in the future – or whether it's likely to annoy some people so much that they'll push for laws to stop it. So a group of us were asked by the Nuffield Bioethics Council to think about what happens when cloud-based medical records collide with pervasive genomics and social data generally. Our report, due out February 6th, discusses not just what medical ethics might look like in the age of big data, but the more general question of how to tackle the ethics of building a new system that may affect everyone. We propose a fourfold way: respect for persons; respect for human-rights law; dialogue with those who have a morally relevant interest; and effective transparency mechanisms.

Please see the link for more information about this report - https://www.lightbluetouchpaper.org/2015/02/03/nuffield-bioethics-report/

Big Data - A Big Issue for Official Statistics

Jane **Naylor**, Senior Principal Methodologist, Office for National Statistics

The Office for National Statistics (ONS) is the UK's largest independent producer of official statistics. It is responsible for collecting and publishing statistics related to the economy, population and society at national, regional and local levels. These statistics are used within central and local government to underpin policy making and to plan services. The ONS recognises the importance of examining the potential of big data sources and related technologies. This presentation will cover progress being made at the ONS to investigate the potential advantages of using big data, to understand the challenges with using these data sources and to establish a longer term strategy for big data within official statistics to support policy making. Particular focus will be given to case studies from ONS and other International Statistics Organisations and partnership working with key stakeholders such as academics and commercial organisations. We believe big data is a big issue for official statistics.

A manifesto for data informed policy

Hetan **Shah**, Executive Director, Royal Statistical Society (RSS)

Hetan will talk about the Royal Statistical Society's Data Manifesto which was launched in September 2014 and outlined what the next Government should do to ensure the maximum impact of data on policy. Hetan will argue that 'big data' is a loose term which is better understood as the ubiquity of data, be those datasets large or small. He will argue that there are considerable opportunities for greater evidence informed policy, including through the recently developed What Works Centres. He will suggest that data sharing within government for statistics and research presents both opportunities and challenges. He will discuss the skills that policymakers and politicians themselves have to make use of the data that is available. He will describe recent Royal Statistical Society / Ipsos MORI research indicating that there is a 'data trust deficit' amongst UK institutions, and will put forward a number of suggestions around strengthening trust, including ending the practice of pre-release access of statistics.

Hetan will talk about the opportunities around engaging local communities to local datasets. He will discuss the role of the private sector, including what kind of standards they should be held to in the data they publish and hold. He will also consider the 'open data' agenda and assess where it has got to in the UK and what more needs to be done. He will make the case for investment in the statistical and data system in the UK in order to utilise new kinds of data and to innovate. He will also consider the wider educational needs posed by the 'data economy' and suggest what changes are required to help the UK grasp the opportunities before it.

Big data: big opportunity; big brother; or big trouble?

Jo **Dally**, Deputy Director for Data Analysis, Horizon Scanning and Project Development, Government Office for Science (GO-Science)

The world is being shaped by the digital revolution: the internet of things is upon us; Moore's law has jumped from transistors to data; and the UK government itself is fuelling the information economy as it pursues an enthusiastic open data agenda. A massive increase in data availability is complemented by ever-increasing sophistication of algorithms used to interpret it. This enables us to derive enormous utility from the data – to buy, sell, communicate, do research, or navigate. But as we engage with the digital world, we are giving out a huge amount of information that other people can use, and our individually held identities are increasingly more difficult to distinguish. What does the future hold for identity, identification, privacy and autonomy, and what role should government play as we navigate the risks and opportunities of the age of big data?

(Title TBC)

Antony **Walker**, Deputy CEO, techUK

Genomics England and delivering the Prime Minister's 100,000 genomes initiative

Mark **Bale**, Deputy Head of Health Science and Bioethics Division at the Department of Health

Reconciling the interests of individuals and populations in policy development in genomics

Alison **Hall**, Public Health Genomics Foundation

For decades medical research and health services have generated genetic data, in the knowledge that it can be predictive of future health. This testing has now expanded beyond the chemical DNA, or genes – the inherited elements that code for protein development that are found in each of our cells. Genomic tests extend beyond DNA to cellular proteins and the cellular environment: these have been developed to look, in increased detail, across the entire genome, enabling comprehensive assessments of current and future ill-health. Whilst these recent developments in genomic sequencing technologies have the potential to revolutionise the diagnosis and treatment of many diseases, particularly inherited diseases and cancers, the characteristics of the data generated by these technologies raise unprecedented challenges for genomics researchers, health providers and policy makers. The clinical utility of genomic sequence data can only be realised through having sufficient infrastructure and expertise in place to be able to store, analyse and interpret this data. This process of interpretation is not straightforward and two types of findings are ethically problematic: firstly, genomic sequencing is likely to generate findings that are of uncertain significance. Identifying the thresholds that should be used for reporting and disclosure in a research and a clinical context is a current priority for policy makers, and this presentation explores a number of different approaches. At stake is who should bear the burden of that uncertainty, and how it should be resolved.

A second challenge is where genomic sequencing technologies generate unexpected or secondary findings which nevertheless may be clinically actionable. Both in research and clinical settings, policy makers have tended to address this second class of findings by emphasising the need for transparency and accountability through an enhanced consent process and increased dialogue. As an alternative, some programmes have argued in favour of participants and patients having a right to choose the tests that are carried out and results returned. In some cases this has also included offering parents choices about whether their children should receive predictive genomic testing, or whether comprehensive carrier testing should be carried out to determine the risk of a couple having a child affected by an inherited genetic disease. Some programmes also envisage offering participants unfettered access to raw sequence data for themselves and their children. The paper explores the extent to which these policy developments might exert novel claims of ownership and privacy over this data, the tensions that may arise between individual and population centred approaches as a result, and suggests some broader insights for policy development.

OPTIMISE (Optimisation of Prognosis and Treatment In Multiple Sclerosis) – A platform for maximizing the use of large scale longitudinal multiple sclerosis patient data

May **Yong**, Paul **Matthews** and Yike **Guo**, Imperial College London

Multiple Sclerosis (MS) affects over 100,000 people in the United Kingdom. It is an idiopathic inflammation disorder of the central nervous system. The natural history of MS is highly variable, and the course of this disease is highly heterogeneous.

Existing studies have shown that there are two major features relevant for stratifying patients to treatments. These factors are disease severity estimation and early treatment response to rapidly identify those patients who require additional treatment.

The integration of multiple data types recorded longitudinally from patients is necessary so that biomarkers can be discovered and algorithms be developed to stratify patients in order to maximize treatment efficacy.

A patient registry that can be used for MS research purposes has to take into account multiple data types, from commonplace clinical data to novel data types such as gait or new functional tests and questionnaires which is being developed in current research.

The registry needs widespread support from both clinicians and researchers, in order to achieve the number of subjects necessary for generating statistically significant results. These data types have to be recorded in a standardized manner so that data from multiple sites can be integrated and compared. The vocabulary for describing data has to be controlled so that queries can be made across federated data sets.

In addition, a standard set of analytical tools have to be made accessible so that cohorts can be selected from the registry data. Finally, the datasets along with their analysis results have to be made accessible so that there is transparency about results reproducibility, and updated patient data that changes analysis results can be tracked and investigated.

OPTIMISE addresses these requirements by linking existing data management and analysis tools, and supplementing them with customised-for-MS applications.

OPTIMISE is designed to act as a large-scale observational registry for MS patients. This project facilitates MS research by bringing together all data types collected from MS patients, ranging from diagnosis and medical condition information, to findings from investigations such as MRI, CSF biomarkers and miRNA data. As a research tool, OPTIMISE also enables the registration of new data types associated with MS such as gait and EEG data.

In addition to its role as an MS patient registry, OPTIMISE offers a set of tools customised to facilitate MS data recording, exploration, organization and archiving.

OPTIMISE enables data integration and maximization of data value by recording clinical data to CDISC data standards. This means that data from existing clinical trials can be compared to data held in OPTIMISE registry. We are also working with MSBASE, an existing MS patient registry to map their data types to CDISC standards, so that MSBASE data that has been collected over the past decade can be integrated with data from OPTIMISE.

We maximize transparency and reproducibility of analysis results by offering data organization, analysis and archiving on a single knowledge management platform. Data is uploaded onto a knowledge management platform called tranSMART, cohorts are selected and analysed in a user-friendly environment.

Therefore, as new data is updated, analysis results such as correlations can be regenerated. Archived data is stored using the same platform so that datasets can be reanalysed without the need to be reassembled and reformatted.

OPTIMISE user flow is as follows: Health care practitioners (HCPs) record patient longitudinal data via our web-based data collection portal. In its most basic form, this tool is built to enable both HCP data input and offers HCPs the capability of tracking individual patient records. The HCP is also provided with a timeline of the patient's condition where each data type is viewed in the context of patient's entire medical history and treatment responses.

In addition to HCP input data, OPTIMISE develops web applications for collecting patient reported outcome measures. This enables to us to build a higher resolution picture of patient wellbeing in terms of mood and fatigue, as patients are able to respond to questionnaires from the comfort of their home. Future work includes integrating eyesight tests, dexterity, GPS –derived data from phones and gait data from body sensors.

All recorded phenotypic and 'omics data are made accessible via a web-based knowledge management analytical platform called tranSMART. Imaging data are stored in a web-based image repository called XNAT. Both XNAT and tranSMART are open sourced applications.

tranSMART enables dynamic cohort selection and provides a set of statistical tools for exploring the assembled cohorts. Users can also download selected cohorts and take them offline for further analysis.

The integration of XNAT and tranSMART enables querying across phenotypical, ‘omics and imaging data. This means that we are able to assemble datasets by selecting parameters across multiple data domains (eg. Select patients with specific MRI brain volume changes, and were treated with particular drugs and whose functional tests show a specific score change).

Data-driven innovation policies in Europe: Mapping methods and sources

Alexander **Kleibrink**, Joint Research Centre, European Commission, Spain

In the European Union (EU), a new multi-billion programme for regional development and innovation will steer economic development policies for the coming seven years. For the first time, it requires national and regional policy makers to design evidence-based innovation strategies based on a comprehensive analyses of socio-economic indicators reflecting strengths and weaknesses of the innovation system. At the same time, policy makers have to develop data-driven monitoring systems to keep track of progress when implementing the strategies. Both exercises imply data challenges for the public sector and the need to outsource parts of the work. This paper provides a first mapping of methodologies and data sources used as an evidence base in regions in five EU countries. It also discusses the on-going efforts to employ new data sources for monitoring innovation activities in these regions. Based on the analysis, I describe the opportunities and the limitations of data mining for economic development intelligence.

Building an index for Sustainable Development Goals Using a Dynamical Systems Data-Science Approach

Viktoria **Spaiser**, Ranganathan **Shyam**, Ranjula Bali **Swain** and David J.T. **Sumpter**, Uppsala University, Sweeden

Defining and measuring development and progress is an important component of development research. The current indices commonly used such as Gross Domestic Product (GDP) or Human Development Index (HDI) proved to be inadequate and alternative indices such as Happy Planet Index (HPI) are currently not backed by sufficient data. In the post 2015 Millennium Development Goals scenario, when the United Nations is working on setting new Sustainable Development Goals (SDG), it becomes imperative to come up with a reasonable index of development that addresses the identified core pillars of development: eradicating extreme poverty, protecting the environment, promoting social inclusion and economic opportunities for all and building peace and effective governance. Likewise the index should be backed by data to be useful. Finally, in contrast to present static indices, it is also important to ensure that the index is constructed such that complex nonlinear dynamics of development progress are captured accurately so we can make reliable predictions on future development progress. The index should not be only a number but a model for development progress. In this project, we combine all these requirements to suggest a methodology (combination of feature selection algorithms

to process around 570 potential predictors, dynamical systems modeling and ensemble learning) to build an index that will be useful in the UN SDG framework.

Innovation Policy-Making in the Big Data Era

Tom **Crick,** Cardiff Metropolitan University; Juan **Mateos-Garcia**, Nesta; Hasan **Bakhshi**, Nesta; Stian **Westlake**, Nesta

There has been an explosion of interest in the potential of big data as a driver of better decision- making in many policy areas, including innovation policy. In this paper, we draw on the theory of innovation, and on Nesta's own experience[1] as an innovation agency in order to address the following research questions:

- *What are the characteristics of innovation policy that make it a suitable domain for the application of big data sources and analytical methodologies?*
- *What is the state of the art, and what are the emerging opportunities for the application of big data for innovation policy?*

In doing this, we seek to provide a firmer conceptual grounding for work in this area, and to set a vision for the development of big data applications addressing the needs of innovation policymakers.

We begin by identifying the main rationales for innovation policy: market failures linked to the fact that innovators often fail to fully capture the benefits of their investments, systems failure caused by gaps in the "system of innovation" that ought to connect innovation agents, and inhibited emergence, where a state of uncertainty about the future configuration of a market or technology field hinders its development [Gustafsson and Autio, 2011].

There are many policy options to remove these barriers to innovation, ranging from direct pub- lic investments on R&D to regulation and procurement [Edler et al., 2013]. Their design, implemen- tation and evaluation has traditionally been based on data sources such as business and innovation surveys, administrative data, and metrics of scientific and technological output (academic publications and patents) [Fagerberg et al., 2006]. Three defining characteristics of innovation do however limit the usefulness of these data sources and outputs for innovation policymaking:

1. **Innovation involves novelty in inputs, processes and outputs:** it is associated with new capabilities, forms of organisation and industries which, by definition, are not captured by exist- ing classifications of economic activity such as Standard Occupational Classifications (SOC) and Standard Industrial Classifications (SIC).
2. **Innovation is not confined to science and technology:** it may reflect changes in, say, business model, marketing or aesthetic: as a result, it is not always captured by traditional metrics such as academic papers and patents.

[1] e.g. http://www.nesta.org.uk/blog/big-data-better-innovation-policy

3. **Innovation is a complex, networked process:** it reflects a dynamic combination of resources and capabilities of many different agents and institutions. Measuring it requires combining data from a multitude of these sources. In turn, those who stand to benefit from access to data on innovation goes beyond policymakers, and encompasses investors, entrepreneurs and corporates, to name a few. However, in practice, most (aggregated, lagging) innovation data outputs are of limited relevance for these agents.

Big data can help overcome some of these challenges for innovation policymaking using conventional data inputs and outputs. Following Schroeder and Cowls [2014], we define big data as datasets of a volume, variety (complexity) and velocity unprecedented in the innovation policy domain, together with new analytical techniques and data outputs (such as data visualisations and interactive data platforms) used to analyse and create value from these data.

Big datasets (e.g. information provided by businesses on their websites) are often unstructured and closer to real-time than official data. This means that they can be used to identify new innovation areas as they emerge, even when these do not respect traditional occupational or industry boundaries. Some metrics that policymakers use to measure innovation are steeped in scientific and technological understandings of innovation; big datasets, by expanding possible sources of data, need not be so con- strained. Big data is high-resolution (when it is based on publicly available data, it is often possible to identify individual agents like businesses or investors in it in ways that official data, which is subject to non-disclosure constraints, is not). This makes it easier to republish it in interactive formats, say, that can be queried and exploited by a variety of innovation agents in addition to policymakers (this is manifest in the recent creation of a variety of online platforms to map innovative industries, clusters and ecosystems using, for example, publicly available data from Companies House).

Big data is however no panacea for policymakers, and its use in innovation policy presents serious methodological challenges. Nevertheless, there exists significant policy work done in the open science space in which to analyse and leverage [Royal Society, 2012]. The innovation landscape is constantly shifting, leading to the arrival of new data sources to study, and structural changes that can impact the reliability of algorithmic data collection and analysis. There is a risk that online data sources might offer a biased representation of innovation activity that privileges digitised industries at the expense of those trading in physical goods and services, and consumer facing industries at the expense of business-to- business sectors. Privacy is of course another critical consideration, especially where personal information is involved [House of Commons Science & Technology Committee, 2014], but there is an imperative for open and sharable data to provide the platform for effective (and transparent) policy-making.

Notwithstanding these important issues, we consider that there is significant scope for innovation in the use of big data for innovation policy – we conclude this paper by outlining some of the main opportunities, and setting out Nesta’s future programme of research and platform development in this space:

1. Go beyond innovation maps based on SOC and SIC coding and official geographies, and make more use of unstructured data collection and (supervised and unsupervised) classification methods.
2. Complement descriptive analyses and sample-based inference with predictive modelling.
3. Explore the opportunities of real-time data and interactive data visualisation for innovation policy.
4. Combine big data sources with official and policy activity data in order to evaluate innovation policy impacts.
5. Develop standards for data-sharing so as to minimise the risk of fragmentation into incompatible platforms capturing disparate aspects of innovation activity.
6. Develop open datasets and transparent (as compared to “black box”) methodologies for big data analysis.
7. Creatively explore the potential use of big data sources and methods for the study of industries which may not currently be data science-literate – and therefore less well catered for by online data – but are of great importance for policymakers, such as manufacturing.

References:

Robin Gustafsson and Erkko Autio. A Failure Trichotomy in Knowledge Exploration and Exploitation. Research Policy, 40(6):819–831, 2011.

Jakob Edler, Paul Cunningham, Abdullah Go˙k, and Philip Shapira. Impacts of Innovation Policy: Synthesis and Conclusion. 2013. Compendium of Evidence on the Effectiveness of Innovation Policy Intervention Project, Manchester Institute of Innovation Research, University of Manchester.

Jan Fagerberg, David C. Mowery, and Richard R. Nelson, editors. The Oxford Handbook of Innovation. Oxford University Press, 2006.

Ralph Schroeder and Josh Cowls. Big Data, Ethics, and the Social Implications of Knowledge Production. Paper presented at Data Ethics Workshop, KDD@Bloomberg, 2014.

Royal Society. Science as an open enterprise. 2012.

House of Commons Science & Technology Committee. Responsible Use of Data. 2014. Fourth Report of Session 2014-2015.

Shaping Responsible Gambling Policy: a case study of harm minimisation research

Dave **Excell**, Featurespace; Georgiy **Bobashev,** RTI; Heather **Wardle**, NatCen; Daniel **Gonzalez-Ordonez,** Featurespace; Tom **Whitehead**, Featurespace; Robert **Morris**, RTI; Paul **Ruddle,** RTI; Amelia **Caron,** Featurespace

This case study which covers all aspects of Data for Policy's areas of focus: information and evidence in the digital age; policy-making mechanisms and modelling approaches; existing methodologies and best practices for use of Big Data in policy; data collection, storage, processing, and access procedures; cumulative learning in digital environments, potentials in policy context, challenges, and limitations; interaction of domain expertise with digital processing technologies, dealing with imperfect/uncertain data, and the psychology or behaviour of decision-making; security and privacy issues and their intersection with law and ethics.

Recently, UK government has been tasked with responding to public outcry over the addictive nature of Fixed-Odds Betting Terminals (FOBTs). Increased scrutiny from media and lobby groups such as the Campaign for Fairer Gambling put pressure on Government to make FOBTs illegal or put a cap on use, both of which have serious commercial repercussions for the gambling industry.

The Department of Culture, Media, and Sport requested policy recommendations to curb problem gambling, but insufficient academic evidence existed—either because the study was too small, did not include a representative sample size, or was not sufficiently detailed for policy decisions to be made. The DCMS requested an independent, research-based approach to problem gambling before making a policy decision. The Responsible Gambling Strategy Board asked Responsible Gambling Trust to construct a consortium of researchers to provide answers to key questions: whether industry-held data from FOBT machines could be used to identify harmful patterns of play, and if so, to draw implications for responsible gambling interventions.

This package of research required a range of projects to be conducted-- from mapping the theoretical markers of harm and exploring the extent to which the markers are evident within industry-held data, to surveying loyalty card customers and matching responses with industry-held transactional data, to exploring views towards loyalty cards and consumer interventions. In anticipation of the results, Prime Minister David Cameron cautioned that government should wait for the research to be published in December 2014 before taking any action as the programme was expected to provide a methodological template for best practices in Big Data analysis for gambling policy.

We detail the process through which such answers were found, providing a road map of crafting public policy. From academic research to analysis to results, this research is instrumental to crafting a responsible gambling policy that will create a safe and enjoyable environment, addressing the needs and interests of all stakeholders.

This ground-breaking research project represented a number of world firsts, with wide-reaching policy implications. This was the first time the five largest bookmakers in Great Britain made their data available for analysis by independent researchers, as well as the first time that land-based industry data have been screened and analysed alongside information from a problem gambling screen. No large-scale independent research project into problem gambling behaviour had ever been conducted in Great Britain.

Methodology:
Theoretical markers of harm were scoped from existing literature and reviewed to ascertain if they were measureable. Industry data was examined against the agreed measures to ensure a significant statistical distribution. Next, surveys were conducted to measure problem gambling as a proxy for harmful play, using a widely accepted screen: the Problem Gambling Severity Index (PGSI), which produces a problem gambling risk score on the basis of survey question responses.

(PGSI) scores were obtained from players who held FOBT loyalty cards with consideration toward privacy, as participants were required to consent to having their loyalty card data linked to their survey responses for analysis. 4,001 participants agreed to the linkage—making this the largest problem gambling research programme ever conducted in Great Britain.

Two models were built to predict problem gambling harm, against a baseline measurement: the Association of British Bookmakers (ABB) Code for Responsible Gambling. The ‘player model’ was based on behavioural analyses in loyalty card holder data, and provided a 66% improvement in accuracy in detecting problem gamblers. The ‘session model’ was based on proxy measurements for anonymous players, and provided a 550% improvement over the ABB measurement.

Data used:
Industry data was supplied by the five major Licensed Betting Offices in the UK (Betfred, Coral, Ladbrokes, Paddy Power, and William Hill), and their gaming machines suppliers (Inspired Gaming and Scientific Games). The data covered 10 months, from 1 September 2013 to 30 June 2014, and no significant data quality issues were identified that would invalidate results.

Researchers faced unique challenges when designing the methodology. Firstly a large data volume--just under 10 billion data records were provided for the analysis, requiring consideration of how to store, access, and process it efficiently and accurately. Nonetheless, event variables were limited. Data skewedness was also a concern, as was the representativeness of loyalty card players. These challenge provide insight into data collection, storage, processing, and accessing procedures in the context of policy making.

Overall, data related to:
333,091 uniquely identifiable customers
8,289 unique shops
32,650 unique Gaming Machines
9,550,448,367 analysed machine events, including 6,768,053,704 bets
661 different games

Specifically for surveyed customers, data related to:
3,988 loyalty card customers
4,374 unique shops
524,277 gaming machine sessions
35,668,298 bets placed

Key findings:

It is possible to distinguish between problem gamblers and non-problem gamblers in industry data. The key findings are that for the player model, accuracy was 66% greater than the baseline model, and the session model was 550% more accurate than the ABB baseline. However, the ABB baseline was only slightly better than random.

When ‘at-risk’ problem gamblers are removed from analysis, false positive rates can be significantly improved, indicating that many of the false positives are likely to be ‘at-risk’ gamblers.

The players studied in this research are more engaged, and therefore results are likely to represent conservative estimates as regards the accuracy of distinguishing problem and non-problem players.

In contrast to suggestions involving a FOBT stake cap, the research demonstrates that it is not possible to accurately identify problem gamblers through a single variable but requires consideration of a combination of variables.

For example, the research revealed that time of day impacted gambling risk, combined with stake size and type of game. Average stake size doubled after 10 pm, likely related to preference for higher-stakes games late at night. It’s unclear why these patterns are occurring, suggesting those gambling at that time may be at greater risk of harm.

The package of research provides a mechanism to improve measurement of problem gambling, and to translate industry terms like ‘chasing losses’ into quantifiable operational terms for future analysis. We have reached a better understanding of problem gambling, and created template for future large-scale research into social health issues for policy creation.

This finding is crucial for best practices in the use of Big Data in policy, and in establishing a methodology for a more in-depth future study where none existed before.

Big Data is Good for Big Bad Policy: Why big data cannot help complex social policy and planning

Emma **Uprichard**, University of Warwick

There is a growing hope - or rather hype - that big data is going to help craft better policies, faster planning systems, and more robust evidence. We see this across a wide range of areas, including cities and urban planning, health systems and social care, education and pedagogy, and so on. This presentation will provide a critique of the view that big data is going to help solve some our most complex issues. This is not to attack the view that big data can be beneficial for policy purposes, but rather the aim is to outline key reasons why big data cannot, sadly, respond adequately to some of the world's most pressing problems. Drawing on the view that social systems are essentially complex systems, which are dynamic, nonlinear, involving multi-dimensional feedback and feedforward loops over time and space, the paper proposes five key issues that fundamentally limit the hope – and question the rational of the hype – underpinning the view that big data can successfully be used for policy planning. The five issues are the following.

First, methodologically, our big data mining practices fail to produce evidence that is adequate at the level of cause of meaning, which is fundamental to policy planning and practice.

Second, epistemologically, big data cannot empirically capture the range of social mechanisms that are needed to understand to be able to even begin producing evidence based policy planning and practice;

Third, ontologically, big data are themselves part and parcel of the social systems that seek to model. In turn, therefore, the notion that they can be used to know the social world (more) objectively is defunct. Since there is always a recursive interaction between i) how and why big data are produced and by whom, ii) the empirical descriptions that can and are produced with big data, and iii) the decisions and decision making processes based on the available data and the descriptions and explanations derived from it, there are also always going to be limits to how useful big data can actually be with the policy arena.

Fourth, temporally, big data offers a range of new opportunities to model social systems in real-time. Yet, at the heart of policy is the notion that it is possible to produce sustainable notions of the future. There is a discrepancy therefore between the kinds of temporalities embedded within policy planning and practice and those currently at play in big data analytics and datamining.

Fifth, and finally, policy planning ultimately requires an understanding of how things work – how they have worked in the past, what hasn't worked, what has succeeded and importantly, about why these things have happened the way they have happened and not another way. In other words, at the heart of policy planning is the notion of causality and explanation. Yet big data are being use without any theoretical notion of social change and continuity. Indeed for the most part, any theory of change that is simply assumed with big data analytics is that social systems work are more or less similarly to those studied in particle physics or ecological models of social insects (e.g. swarm, contagion, aggregate patterns of change etc.). This then begs the question about what it is we are wanting to do with big data and what the implications are if we do turn heavily to big data to address complex social issues.

Overall, then, this paper argues that whilst big data has its place in the future of policy planning, there are limits to what it can be expected to deliver. The paper concludes with some reflections on what the implications what it may mean methodologically to move forward in a big data era and how researchers and practitioners might maximise the benefit of big data for future policy and planning purposes.

Data-driven policy making between myths and reality

Francesco **Mureddu** and David **Osimo**, Universitat Oberta de Catalunya and Open Evidence, Spain

In the last twenty years, we've witnessed the appearance of many "new" approaches to policy making. Evidence-based policy-making has continuously brought forward new methodologies that promised to make policies "scientific", from complex systems approaches, to behavioural studies, to randomized control trials, to network analysis. The open government movement has increasingly moved its focus from open service delivery to open policy-making, with the idea that "total transparency", "open data" and the "wisdom of crowds" could improve policies and reduce the probability of ill-faithed decisions. However, there's little evidence of a substantial improvement in policy-making.

Today new tools such as big data and artificial intelligence promise "automated decision-making" where policies decisions are taken instantly by predictive algorithms fed by terabyte of real-time data, even without the active role of human decision makers. McKinsey estimates the potential savings for European public administrations at around €150 billion to €300 billion a year. Data technology solutions are applicable throughout the different phases of policy cycle, from agenda setting to policy design, implementation and evaluation. To put it very simply, evidence-based policy-making is based on causality, and data-driven approaches appear to offer radical improvements in detecting causality links by cross-analysing very disparate datasets.

But in view of years of failed promises, we should thoroughly assess the concrete opportunities: how do these opportunities play out concretely? What are the risks and the benefits?

Data-driven approaches have recently encountered harsh criticisms. Gartner put "big data" in their "trough of disillusionment". One of the "flagship" initiatives of big data, the Google Flu project, which uses data from searches to predict flu outbreaks, proved ineffective: an article in Science showed that GFT has over-estimated the prevalence of flu for 100 out of the last 108 weeks; it's been wrong since August 2011. Concerns over respect of human rights are always present, and reach beyond privacy issues. For instance, early implementations of data technologies were particularly successful in the domain of security: predpol.com is an application that allows the police force to predict where crime will happen based on a set of contextual data and time series. It helped reducing crime rates substantially, but also came under scrutiny because it ultimately led to pre-identify potential crimes by profiling people, neighbourhoods and behaviours.

The paper presents 4 real life cases of application of big data to public decision making to different phases of the policy cycle. The analysis shows that open government data are an important enabler, but the impact is reached when (non-open) administrative data are analysed and contextualized with open government data. The impact can't be exclusively attributed to big data solutions, but is delivered in with innovations in service delivery models. To overcome concerns, we need to provide the right expectations and not present big data as a magic bullet. Effective data-driven approaches come from the combination of theoretical models and data, and from traditional “small” data and “big” data. To take advantage of these opportunities, the continuous, iterative and in depth and collaborative work between policy analysts and data scientists is needed.

The paper illustrates the concrete, real-life benefits of a data-driven approach for policy-makers that goes beyond the cycle of “hype” and “disillusionment”, and to enable sustainable innovation in policy-making.

Mining Data about the Public as a Tool for Policy

Ralph **Schroeder**, Josh **Cowls** and Eric T **Meyer**, Oxford Internet Institute, University of Oxford

Big data is nowadays seen as an important tool for policymakers, but so far there have been few analyses of its efficacy. This paper focuses on mining data about the public’s opinions, preferences and behaviours as a tool for public policy decision making. A number of tools and areas will be examined, all with advantages and limitations.

First, there has been extensive debate about Google Flu Trends (Lazer, Kennedy, King, and Vespignani 2014), including problematic methodological issues as well as issues to do with access to the data and therefore about replicability. It has been shown, however, that Wikipedia use can also be used as a tool for predicting flu and other diseases, without issues about data access and replicability (Generous et al 2014). The use of Wikipedia nevertheless raises a number of issues for policymaking, including which data sources are best suited for decision-making, given that Wikipedia is not used everywhere to the same extent (West, Weber and Castillo 2012).

A second example is gauging public interest by means of search engine behaviour in climate change. Here Anderegg and Goldsmith (2014) have shown, using Google Trends, how public interest has waxed and waned, and different search terms (‘global warming’ versus ‘climate change’) can give indications of what the public thinks. Further, it is possible to identify climate change sceptics (those who search for ‘global warming hoax’), and whether these sceptics correspond, for example, to places where such scepticism might be expected (such as Republican-leaning states in the United States). For policymaking, it is important in this case to have an understanding of the extent to which Americans and other use the Internet to find information about science, which has been a majority for some time (Horrigan 2006).

A third area is economic policymaking. The ‘Billion prices’ project (Cavallo, Cavallo and Rigobon 2014), for example, has related price fluctuations to natural disasters. Prices that are scraped from the web can also be used to complement or provide more timely and accurate indications of economic trends (Cavallo 2013). Here there may be trade-offs between the richness of traditional data sources data, which are also embedded in context, as against the ease and timeliness of online data.

Fourth, there are mobile phone records for tracking populations after an earthquake disaster (Bengtsson et al. 2011), which has provided accurate information in support of relief operations. This type of effort depends on collaboration with mobile phone operators for access to the data.

Fifth, microblogging can be used to gain insights into political issues that the public are interested in (Neumann et al.2014) and what kinds of spreads of messages are most likely to lead to success in political activism (Fu and Chau 2014). Again, biases in the populations using these tools are an issue (Gonzalez-Bailon et al. 2014). This kind of research could be put to positive uses by activists, but the research could equally be used by authoritarian regimes such as China to nip the spread of messages in the bud.

In all these cases, there are a number of limitations for policymaking: one is that the relations between the data sources used (search engines, Wikipedia, mobile phone records) and the phenomena under investigation skews the actionable knowledge towards those about whom the data are captured. This is not just a question of the representativeness of users, but also potentially opens up a divide between those with access to the technology as against those who do not (Lerner 2013). The second is that policymakers' needs are likely to be highly specific, while big data research is oriented towards wider generalizability. A third limitation is that the chain from big data insights to policy implementation – as with other knowledge informing public policy – can be more or less direct.

One of the advantages of big data for policymaking is that it is more resource-effective and timely than other means. It can also provide a more disintermediated picture of public attitudes and behaviours, which are often filtered through the media or self-report and the like. On the other hand, this method can lead to more intermediation, since digital traces are always a representation of peoples' attitudes and behaviour compared with direct observation, including of those who do not leave digital traces. Thus new methods may not be as accurate as older methods.

These advantages and limitations can be related to the extent to which policymaking depends on the coupling between the insights from big data analytics and the populations or social environments to which they are applied: Unless there are feedback loops between these three elements (policymaking, insights from data, and the populations that the data pertain to), big data analytics are unlikely to be useful. However, these feedback loops require further distinctions: between planning, forecasting, and routine access to information. These three policymaking activities have different needs for big data: planning depends on pulling together information for a goal; forecasting is particularly dependent on models and simulations; and routine access requires the right resources to be ready-to-hand. These three activities are often combined, but the reason that big data analytics have not become more prominent is that few of these three activities have so far gone beyond pilot projects.

A broader criticism of policymaking that uses scientific evidence is that this scientization falsely lends an aura of objectivity to what is inherently a political process. However, data-driven techniques, if they are based on data as defined here (in the full version of the paper), are bound to turn policymaking not just into a scientific, but also a machine-like process, guiding this process with external evidence which is computationally driven. The benefits (a more powerful process) and drawbacks (the costs of greater control) are evident. However, both benefits and drawbacks are currently exaggerated since the cases described here are far from being operationalized among policymakers (unlike among researchers, and in the private sector), and this is because of the loose coupling between knowledge in the policymaking process. Nonetheless, these obstacles are bound to be shortly overcome, making greater understanding in this area all the more urgent.

References

Anderegg, William R. L., Gregory R. Goldsmith. 2014. Public interest in climate change over the past decade and the effects of the ‘climategate’ media event. Environmental Research Letters 9 054005

Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J (2011) Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. PLoS Med 8(8): e1001083. doi:10.1371/journal.pmed.1001083

Cavallo, A. 2013. "Scraped Data and Sticky Prices" - MIT Sloan Working Paper No. 4976-12.

Cavallo, A., Cavallo, E., Rigobon, R. 2014. "Prices and Supply Disruptions during Natural Disasters" (with Eduardo Cavallo and Roberto Rigobon). Review of Income and Wealth, Volume 60.

Fu, King-wa; Chau, Michael. 2014. Use of Microblogs in Grassroots Movements in China: Exploring the Role of Online Networking in Agenda Setting, Journal of Information Technology & Politics, 11(3), 309-328.

Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R (2014) Global Disease Monitoring and Forecasting with Wikipedia. PLoS Comput Biol 10(11): e1003892.

González-Bailón S, Wang N, Rivero, A, Borge-Holthoefer J and Moreno Y (2014) Assessing the bias in samples of large online networks. Social Networks 38: 16–27.

Horrigan, J. 2006.The Internet as a Resource for News and Information about Science. Pew Internet Research Project, http://www.pewinternet.org/

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis, Science 343, no. 14 March: 1203-1205.

Lerner, Jonas (2013), Big Data and its Exclusions. 66 Stan. L. Rev. Online 55.

Neuman, W. Russell, Guggenheim, L., Mo Jang, S. and Bae, S. Y. (2014), The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data. Journal of Communication, 64: 193–214.

West, R., Weber, I., Castillo, C. (2012) ‘Drawing a Data-Driven Portrait of Wikipedia Editors’, Proceedings of WikiSym ’12, August 27–29, 2012, Linz, Austria.

Using data to understand how the statute book works

John **Sheridan**, Head of Legislation Services, The National Archives

Day 2 – Tuesday, June 16

Big data: promise and pitfalls

David J **Hand**, Emeritus Professor of Mathematics, Imperial College; Chair, UK's Administrative Data Research Network

In this talk I attempt to take a dispassionate view of the promises and risks arising from so-called 'big data'. I examine the extent to which the concept represents real potential as opposed to mere media hype. I compare the enthusiasts' assertions with those made about data mining some twenty years ago and, indeed, ask what, if anything, is the difference between big data and data mining. Is it mere media rebranding? I give real examples of advances which have been attributed to big data, and ask whether the attribution is justified. I also give examples of where things have gone awfully wrong and try to identify the reasons underlying the failure.

Real-time Policy-making: Merging New and Old Big Data

Mirco **Musolesi**, University College London

The advent of real-time analytics is potentially going to have a strong impact on policy making in governmental bodies at different level (local, national and global), and in decision-making in companies and non-profit organisations. In this talk, I will discuss the current challenges and opportunities in developing models and systems for real-time policy-making with a focus on the integration of new and emerging data sources, such as social media, mobile phone data and so on, with more traditional ones.

Intelligent Web Technology for the Policy-making Process and Policy Evaluation

Eugen **Molnar**, Comenius University and Rastislav **Molnar**, Imperial College London

Policy making process relies on the following key elements - actors, context and events. In the same time actors are specified by various attributes, behaviours and first of all here is communication including publishing opinions in social media or via news networks. It is extremely important to catch relevant feedback and to be able to quickly validate an impact a policy has on citizens. One of the most difficult questions faced by many policy makers today is how to catch a relevant and immediate feedback on new policies in a timely and convenient manner. We aim to answer that question in our presentation.

We propose a framework for information collection and processing from various public sources including social networks and news services. The framework contains two step textual analysis, clustering and classification of textual data and then the concept, entity and information extraction from them. Third step is the knowledge representation, where we utilize the semantic web technology, followed by the data visualization. Our goal is to demonstrate its capabilities and the usefulness in the context of the policy-making process and policy evaluation. We define this framework and present how it can be successfully used in this domain.

Later we present how proposed framework can be used on a concrete use case. Our implementation will utilize existing open source solutions with minor customizations and we plan to present and share this implementation as well. The solution combines Natural Language Processing and rule based system with analytics and visualization tools. We believe rule based systems have an important place in the Big Data field. They represent mature technology and their potential inhere in the power of rules which can be automated and tailored in different domains easily.

This presentation introduces a framework for an automatic information extraction, processing, sentiment analysis and visualization. We will illustrate the power of the rules and possible benefits of using rules based system in the policy making process and policy evaluation. We believe similar frameworks may be adopted and beneficial in other domains like the financial sector, the risk management or for the customer care of private companies as well.

Microdata, Networks, and Simulation for Unemployment Policy

Omar **Guerrero** and Eduardo **Lopez**, University of Oxford

We present an empirically motivated theory of unemployment dynamics that makes use of large-scale employment history records, and the method of random walkers on graphs. The theory explains important stylized facts of labor and firm dynamics (which are usually disconnected in the literature) and provides a solution to the problem of arbitrary partitions of the economy into submarkets (which is key when studying aggregate unemployment). This framework provides analytical insights that are empirically validated with micro-data containing the employment histories of all workers and firms in Finland. Finally, we demonstrate how labor dynamics on networks are susceptible to localized shocks (a feature that cannot be addressed in conventional models). Our research advocates for the collection of fine-grained micro data, the measurement of firm-specific quantities, and the use of computational methods in order to design labor policies that are equipped to deal with a wider range of economic scenarios.

Toward Convergence of Information Theory for Efficient Data Collection, Storage and Access

Konstantinos **Psannis**, University of Macedonia, Greece

Information Theory, published in July and October of 1948 by Shannon, continues to set the stage for the development of communications, data storage and processing, and other information technologies. In the past few years, the boundaries between information technology and communications technology have become progressively identical. Generally, integration of information technology and communications technology produce the convergence of Information and Communications

Technology (ICT). One of the main objectives of ICT convergence is to increase interoperability between services and applications for industrial and social innovation. This paper proposes the convergence of the developments of Information Theory in order to increase interoperability for efficient data collection, storage and access. This paper attempts to explore this area and classifying the convergence of Information Theory developments benefits. The benefits have been classified according to the current priorities policy areas in the field of data analysis. In doing so, the paper allows researchers, managers, developers and consultants to better understand the strength of Information Theory convergence for efficient data collection, storage and access.

Big data for evidence-informed policy: International inventory and European Commission case studies

Eric **Meyer**, Oxford Internet Institute (OII)
Prabhat **Agarwal**, Head of Sector ‘Evidence-based policy-making’, DG CONNECT, European Commission
Martijn **Poel**, Technopolis Group
Ralph **Schroeder**, Professor, Oxford Internet Institute, University of Oxford

Big data can be defined as a step change in the possibilities to collect, link, analyse and visualise data. Data sources include social media, real-time sensor data, public administration data (including open data), data from statistical offices, commercially traded data, and the like. The increased volume and variety of data, together with advances in connectivity, data storage and computing, allow for improvements in data analytics. Options for visualisation of data go well beyond interactive maps, heat maps, timelines and network graphs.

Recent studies have explored how relevant technologies and services are being developed and commercialised by researchers, software firms, data analysts, web designers, etc. Along the same lines, recent studies provide indications about the benefits of big data for lead users such as retail and insurance.

Much less is known about the opportunities and use of big data by public policymakers. Policymakers have started to explore how big data can be used throughout the policy cycle of agenda setting, policy design, ex ante impact assessment, monitoring and ex post evaluation and impact assessment. One of the challenges is to ensure that well-designed indicators, more data and new analytical tools actually lead to better insights and policy impact. Another challenge is to address privacy and other legal and ethical issues.
The knowledge gaps mostly concerns national and international initiatives, as regional and local initiatives required less coordination and resources and several cases have been documented in case studies (e.g. on crime, traffic or housing prices).

The European Commission has commissioned Technopolis Group, the Oxford Internet Institute (OII) and the Centre for European Policy Studies (CEPS) to conduct an inventory of national and international initiatives that use big data to inform policy making. The topic cuts across all policy areas and societal challenges, therefore the study will not be limited to the portfolio of the Commission's Directorate-General for Communications Networks, Content and Technology (the client), but will include other European Commission services and policy areas.

The study will produce:

1. An inventory of big data for policy initiatives in European Union Member States and in Canada, India, Singapore, South Korea and the US. Further, relevant initiatives at the international level (EC, OECD, WHO, UN, etc.) also fall within scope. The inventory will address, for each initiative:

    - Policy area
    - Policy level
    - Responsible public authority o Data sources used
    - Data linking
    - Data analytics tools
    - Policy theory/framework model that underlies the analysis o Relevance for different parts of the policy cycle
    - Visualisation tools
    - Addressing privacy
    - Addressing inclusion and other ethical and legal matters

2. A report about state-of-the-art in big data for evidence-based policy, building on the inventory, a literature review, stakeholder mapping and interviews with thought leaders.

3. Six case studies (use cases) for the European Commission, in collaboration with the relevant EC services. The case studies will address policy areas in which big data provides substantial opportunities to improve policy analysis and policy design, given the availability of data, possibilities to link data, shortcomings of existing analyses, etc.

4. One of the six case studies will be developed into an online demonstrator.

5. An international workshop with thought leaders and practitioners will be held in September 2015 in Brussels.

Throughout the study, experts, stakeholders and other interested parties will be invited to contribute with their insights and to give feedback on the main findings, both online and offline.

April 2015, the results of the first three steps of the study will become available to experts, stakeholders and other parties engaged in the study. Thus, the timing of the Cambridge Conference ‘Policy-Making in the Big Data Era’ (15-17 June 2015) is perfect to share these results with a wider audience.

This is a slightly different type of session from the main types outlined for the conference, but we feel that the findings from this project are directly relevant to the theme of this conference, and will be of broad interest to attendees. In order to maximise complementarity to other contributions to the conference, we propose to focus particularly on the inventory of national and international initiatives (and the analysis across initiatives) and the European Commission case studies.

Transforming healthcare data through analytics

Mike **Standing**, Deloitte

Health care systems accumulate significant amounts of clinical, operational and financial data in their everyday operations. This information can be repurposed to provide insights to drive service redesign to improve outcomes and quality.

Mike Standing will introduce an approach based on five principles which highlight the most effective way to use health care information to improve outcomes, quality and increased productivity. Throughout his presentation he will refer to case studies across multiple therapeutic areas including cancer and diabetes.

Data mining to leverage change in clinical practice

Oriol **Sola-Morales** and Eduard **Gil**, Sagessa, Spain

Sagessa is a vertically and horizontally integrated healthcare organisation.

Despite using diferent front end interfaces, we have been able to track patient data in different settings and from different origins to create a master file that would allow us to track patient flow and management of certain diseases.

We shall present as case studies how we have been tracking data of patients in diferent settings having had a stroke event and Type 2 Diabetis to ensure quality of medical practice and to further impement changes in patient management.

Assessing the Real-World Data Policy Landscape for Health and Healthcare in Europe

Joanna **Chataway**, RAND Europe

In 2014, RAND Europe, supported by IBM, carried out a study to assess the European policy environment for real world data (RWD) related to health and healthcare.

The research detailed current forms and uses of RWD in Europe and highlighted their significant potential for assessing the (short- or long-term) impact of different drugs or medical treatments and for informing and improving healthcare service delivery. Although the potential of RWD use seems quite clear, this research reveals barriers that restrict further development towards its full exploitation:

- the absence of common standards for defining the content and quality of RWD
- methodological barriers that may limit the potential benefits of RWD analysis

- governance issues underlying the absence of standards for collaboration between stakeholders
- privacy concerns and binding data protection legislation which can be seen to restrict access and use of data.

The study revealed a variety of ways that constraints are being overcome at national and European level. The research concluded that to maximise the potential of these new pools of data in the healthcare sector, stakeholders need to identify pathways and processes which will allow them to efficiently access and use RWD to achieve better research outcomes and improved healthcare delivery.

Clinical genomic data sharing for healthcare: practical and policy challenges

Chris **Rands**, Public Health Genomics Foundation

Genomics is broadly the study of the genetic material of an organism, and increasingly a discipline concerned with developing and deploying approaches for mining large and complex datasets to identify trends, correlations, and patterns. In healthcare, genomics based approaches have the potential to improve and inform the understanding, diagnosis and management of conditions where there is some underlying genetic basis, such as inherited diseases and cancers. The envisaged benefits of genomics in medicine include but are not limited to; faster and more accurate diagnosis, more personalised treatment based on individual's genetics, and new medical and public health interventions underpinned by greater understanding of the relationship between genes and health. A key challenge in realising these benefits is the ability to analyse and interpret the complex and voluminous data generated by genomic investigations. For example a single human genome contains over 3 billion data points, and the biological or clinical significance of the majority of the genome is not yet fully understood.

Aggregating and exchanging genomic and clinical data is an important component of realising and maximising the potential benefits of 'big data' in genomics. Firstly, the analysis of genomes requires access to pre-existing knowledge and genomic data. For example, when determining which part of a patients genome are responsible for their disease, comparison with data from unaffected individuals as well data from patients with the same condition is needed to exclude or include genetic differences as potentially disease causing. Secondly, only by pooling genomic and associated clinical data is it possible to identify relationships between genes or other parts of the genome with clinical traits, and by doing so advancing the application of genomics in medicine.

However, the collation, aggregation and sharing of genomic data is not without challenges, which include unresolved ethical, legal and political considerations around data privacy, ownership, and intellectual property rights, as well as practical and technical hurdles to devising and maintaining infrastructure to share large quantities of data. We explore these challenges in the context of genomic data sharing in healthcare networks, and present policy positions that may facilitate the effective and responsible sharing of data. Many of the challenges around data sharing in genomics may also resonate across other disciplines that utilise ‘big data’, particularly in the healthcare context.

The Metric Tide: big data and the future of research assessment

James **Wilsdon**, Professor of Science and Democracy at SPRU (Science & Technology Policy Research), Sussex University

Citations, journal impact factors, H-indices, even tweets and Facebook likes – there are no end of quantitative measures that can now be used to assess the quality and wider impacts of research. But how robust and reliable are such indicators, and what weight – if any – should we give them in the management of the UK’s research system? Over the past year, the Independent Review of the Role of Metrics in Research Assessment and Management has looked in detail at these questions. The review has explored the use of metrics across the full range of academic disciplines, and assessed their potential contribution to processes of research assessment like the REF. It has looked at how universities themselves use metrics, at the rise of league tables and rankings, at the relationship between metrics and issues of equality and diversity, and at the potential for ‘gaming’ that can arise from the use of particular indicators in the funding system. The review’s final report, The Metric Tide, will be published in July 2015. In this talk, James Wilsdon, chair of the metrics review, will preview its findings and link these to broader debates about the potential and pitfalls of big data.

Data science in government - the benefits and challenges of implementing new analytical techniques and technologies in government

Sue **Bateman**, Head of Data Science, Cabinet Office

The last few years have seen significant advances in the capability and cost of technology, which in turn has led to advances in data science. As the government improves public service productivity and its services become increasingly digital, the opportunities for data science are clear. The Cabinet Office, data scientists in the Government Digital Service, the Government Office for Science and the Office for National Statistics have been working as a joint team to explore these opportunities and also what challenges they bring to government.

This presentation will give a brief overview of the opportunities and challenges of data science as well as explaining the progress we have made so far.
Data science comprises many different techniques, but over the course of our work, certain themes have arisen which have particular benefit to government.

- At a fundamental level, simply presenting data in a more dynamic, interactive way can allow both analysts and policy makers to gain insight in data that they may not have seen in thousands or millions of rows on a spreadsheet. Several of the alpha projects we have undertaken have been successful, not only because of the underlying analysis, but also because the data is now accessible and easier to explore.

- Data science allows us to access data sets which we would couldn't previously analyse, such as social media data, letters or phone calls. This can help us to improve our understanding of citizen and customer views to help with policy development and improve services.
- Machine learning is one of the fundamental techniques that sets data science apart. With more services becoming digitised and real-time flows on information a reality, we can use predictive techniques to anticipate change and resolve service issues more quickly.
- Another of the most sophisticated data science techniques is segmentation. We can use segmentation to tailor services for citizens to find a service that works for them and reduce waste.

The work of the data science team has focused on practical trialling and learning through short, demonstration projects which we will show later in this session. Through these projects and extensive engagement both inside and outside government, we have explored what existing capability there is in government and how we can help departments embed a data science approach to policy making and operations. We have also discovered that technology and data are key challenges for analytical teams doing data science and we are helping to improve the situation both through better IT capability and access to data.

Using the National Energy Efficiency Data Framework (NEED) to better target energy efficiency measures

Shahzia **Holtom**, Data Scientist, Government Digital Service

The Department of Energy and Climate Change (DECC) is working to improve the energy efficiency of the housing stock, including the implementation of schemes to provide energy efficiency measures to homes, such as cavity wall insulation and loft insulation.

DECC also put together the National Energy Efficiency Data-Framework (NEED) which combines gas and electricity consumption data, information on energy efficiency measures installed in homes, property attributes and household characteristics.

We started by combining the NEED data with housing benefit data from DWP - this enriched the information we had, but also gave an opportunity to prove what could be done with new data and new techniques. We then built a predictive model that highlights priority areas which identify those locations where there are households receiving benefit and in need of insulation.

To enhance this analysis, we used one of the more advanced techniques that data science offers; machine learning. In this case, we focussed on one energy efficiency measure, cavity wall insulation, and were able to define certain characteristics of a property that can be used to predict whether it has cavity wall insulation or not – e.g. property age, floor area, loft insulation.

The NEED data set identifies around 7% of properties with cavity wall insulation. A Random Forest machine learning algorithm (a decision tree technique that can handle multiple variables) was used to identify which characteristics of a property could be used to predict whether a property does or does not have cavity wall insulation. The analysis found that factors such as property age, floor area and presence of loft insulation were key factors in predicting cavity wall insulation in a property (consumption data was not included as this could skew the results for low energy users).

These results were then combined with the initial analysis of gas consumption and housing benefit and areas were given a score which determines how high priority they are for being targeted by energy companies for the installation of energy efficiency measures. The images below show how the model has narrowed down the priority areas significantly - image 1 has several highlighted areas which could be classed as a priority whereas as image 2 narrows this down to only the highest priority (darker colours indicate higher priority).

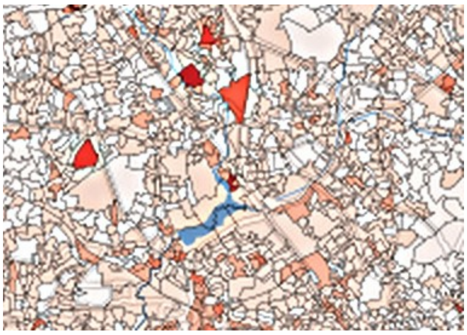


Image 1: priority areas using previous model

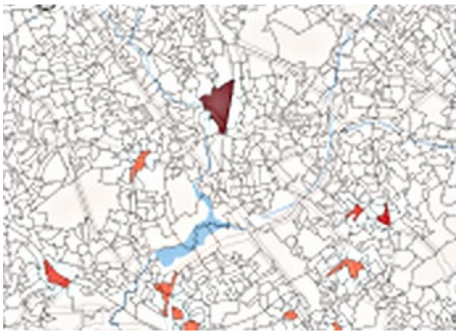


Image 2: priority areas using new model

The results of a project such as this can have an actual impact on households who currently struggle with energy bills or cannot heat their houses fully due to high costs. By providing insulation or draft proofing where it is most needed, we can help to reduce their energy costs and keep houses warm.

Exploring how complaints from departmental and non-departmental sources can predict surges in service demand

Dan **Heron**, Data Scientist, Government Digital Service

This presentation will be a live demonstration and explanation of the methodology. This project has been designed to analyse unstructured complaints data from a range of sources for:

- Government Service managers and stakeholders to understand public perception of their service;
- Service managers to respond quickly to arising problems;
- Designated complaints handlers to sort and prioritise complaints quickly.

This project utilises a variety of analytical tools applied to layered data from multiple sources to visualise a “mean shift” (a significant spike in activity around topics). The system includes historical sliders to be used and for snapshots of activity over specified time parameters to be viewed. This can be done ‘per service’ using a layered time series. If a breakout time series happens then it is sufficient to indicate an issue to investigate.

The system uses a topic model for correspondence data to see what core themes are prevalent across a variety of sources. Using “shiny” (a package in R) to visualise recurring words and phrases has allowed exploration of the topic models — once these are assigned they can be looked at as a time series, with a focus on transactional services. Once a dashboard environment has been produced for one topic area, the system can then be scaled to include many more.

Using web scraped data to supplement the Consumer and Retail Price Indexes

Nigel **Swier**, Principal Methodologist, Big Data Project, Office for National Statistics

The Office for National Statistics (ONS) have a Big Data team who have been working to investigate the potential benefits and challenges of using big data and its associated technologies for official statistics. A key aim of this 15 month project is to develop a strategy for big data for ONS - this will done in part by completing pilot projects to test some big data techniques and new data sources.
One of these pilots involved scraping data from web pages of supermarkets and using that data to supplement the Consumer Price Index (CPI) and Retail Price Index (RPI). This would offer a range of potential benefits over physically visiting stores including reduced data collection costs, increased coverage (more basket items/products) and increased frequency of collection. Some of the key issues to explore in this pilot are:

- the technical feasibility of using web scraped data;
- methods for quality assuring web scraped data;
- comparison with current methods and exploring methodological issues;
- the costs and benefits of scraping web data.

The scrape was set up to get data on 40 items from 3 supermarkets, on a daily basis. The scrape was also run for two different locations of each supermarket to test whether the results differed.

A major challenge throughout this work has been to link the products across time which can be difficult given the high rate of product turnover or repackaging a product which normally involves issuing a new product code. These issues provide a challenge to linking thousands of products across time in a highly dynamic environment. For this reason, further options are being explored with online comparison websites to use the data they collect as they are already comparing equivalent products over time.

Web scraping has proven to be a cost effective method of data collection with minimal overheads for fixing occasional problems with the scrape, however this cost would increase with the scope of the web scraping. Work is underway to review the methodological implications of using such data and we expect results by March 2015.

The advent of Big Data technologies in Finance

Paul **Jones**, SAS

Recent changes in the technology landscape including Hadoop have enabled private and public organisations to apply massive amounts of analytics powered by parallel computing to large volumes of data in a timely fashion at a reasonable cost. This fundamentals changes some of the dynamics of the market:

- Risk can be assessed and managed much more closely in investment or retail banking
- Pricing can be dynamically altered based on individual circumstances or with the addition of external factors including open source data.

This presentation will cover:

- some of the applications that SAS has been working on with customers, including examples from Retail, Investment Banking and Insurance
- how these new applications could affect the dynamics of the market such as effecting the consumer pricing of financial products
- and will finish by highlighting some of the technologies that are supporting these changes.

Big data and central banks

David **Bholat**, Bank of England

I will cover three points in my brief opening comments. First I will situate big data in the context of the Bank's strategic plan and vision. Second, I will summarise and reflect on major initiatives underway. And third I will link central banks' emergent interest in big data approaches with their broader uptake by other economic actors.

The value of big data and its consequences: ethics, controversies and policy

Carla **Bonina**, Surrey Business School

Big and open data bring excitement about the enormous benefits for the economy, society, and environment and beyond. Novel uses of big data analytics, combined with growing access to new technology and smart devices, is spanning and redefining how companies do business, how public services are provided, and how governments design and implement policies. No doubt, the potential benefits of big data are appealing. However, controversial aspects are usually marginalised in the big data enthusiasm. In this talk I will discuss, from a social science perspective, a number of growing controversies in the realm of big and open data debates. My talk will cover cases and examples that show emerging tensions, such as: big data and identity, open data and privacy, divides between those who profit from big data and those that are marginalised, and the accountability mechanisms of algorithms in policy design. Overall, I aim to reflect about the ethics of big data for business and society at large.

Which data for monitoring financial stability? Policy issues related to collecting, processing, and aggregating increasingly granular financial sector data

Harald **Stieber**, Ralph **Dum** and Prabhat **Agarwal**, European Commission, Belgium

As many Governmental and Regulatory bodies across the world, the European Commission is turning towards data-driven policy-making in several of its domains of competence. In this paper, we outline some of the conceptual, technical and policy issues of using emerging data-related technologies in the context of the European Market Infrastructure Regulation.

Since February and August 2014 data on certain financial derivative contracts began to be reported in line with recent regulatory requirements.[2] These vast amounts of data being reported to be used for the monitoring of financial stability raise a number of conceptual and technical, as well as policy issues.

First, similar to many other data-driven policy problems, the data needs to be brought into a format to arrive at a first population that respects a number of normative criteria, or principles [1, 2]. Data should be internally consistent, the rules that structure the data should be transparent and parsimonious. We provide an overview of the core elements of this basic structure, explaining the concepts of legal entity identifier (LEI), unique product identifier (UPI), and unique transaction identifier (UTI), and how these three dimensions could allow map the ecosystem of financial derivatives markets.

Second, the first layer of such data should maximize the efficiency of methods for data aggregation and visualization techniques [3]. Given the vast amounts of data to be processed, data aggregation and visualization can be considered to be indispensable steps that need to be taken before data can be fit for purpose, fit for policy making. We discuss a number of promising approaches to data aggregation in the context of financial derivatives transactions data.
In a Big Data context the appropriate organization of data is crucial. Ideally, the organizational principles are derived from the inherent structure of the data itself. We discuss a functional algorithmic approach as proposed by the ACTUS project [4], as well as more agnostic approaches to structuring the data exploiting the topological

[2] REGULATION (EU) No 648/2012 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 4 July 2012 on OTC derivatives, central counterparties and trade repositories, L 201, 27.7.2012, pp.1-59.

properties of the data using network theoretical concepts [5, 6]. These methods from the emerging science of networks are equally important for data aggregation, determining underlying relations and helping understand dynamical issues underlying the data.

Third, the computation of indicators for monitoring financial stability, as well as indicators for other policy purposes will need to be complemented by various forms of visualization [3, 7, 8]. We argue that visualization becomes increasingly important the bigger the amount of data even if one can find satisfactory solutions to problems of data aggregation and even if these solutions comprise a satisfactory number of efficient indicators for evidence-based policy making. The required leap of faith to act on abstract information as in the case of numerical indicators simply becomes very large as underlying amounts of data become very large. Visualization can be understood in this context as a strategy to control anxiety in the face of Big Data [9].

However, visualization techniques, as is well known from digital imaging, are special forms of aggregating and presenting underlying information/data.

The aim of the paper is to discuss these issues in the context of financial supervision [10]. We focus on the relative costs and benefits/potential of data collection, processing, aggregation/visualization in this context. Ideally, costs to the user in terms of data collection, its granularity and frequency, the skills required to process and use the data increase in a predictable (not too nonlinear) manner in relation to potential benefits, also by way of comparing some recent concrete examples how to model financial markets segments using micro data.

References:

[1] FSB, Basel Committee on Banking Supervision, *Principles for Effective Risk Data Aggregation and Risk Reporting*. Basel, Switzerland: Bank for International Settlements, January 2013.

[2] Office of Financial Research, *Annual Report*, December 2014, Chapters 5 and 6.

[3] Office of Financial Research, *The Application of Visual Analytics to Financial Stability Monitoring*, OFR Working Paper. 14-02b May 9th, 2014, Revised October 7th, 2014.

[4] http://www.projectactus.org/ACTUS/

[5] Caldarelli G, Cristelli M, Gabrielli A, Pietronero L, Scala A, et al. (2012) A Network Analysis of Countries' Export Flows: Firm Grounds for the Building Blocks of the Economy. PLoS ONE 7(10): e47278. doi:10.1371/journal.pone.0047278

[6] http://www.growthcom.eu/

[7] Sheri Markose, Simone Giansante, Ali Rais Shaghaghi, 'Too interconnected to fail' financial network of US CDS market: Topological fragility and systemic risk, Journal of Economic Behavior & Organization, Volume 83, Issue 3, August 2012, Pages 627-646, ISSN 0167-2681, http://dx.doi.org/10.1016/j.jebo.2012.05.016.

[8] Olli Castrén, Michela Rancan, Macro-Networks: An application to euro area financial accounts, Journal of Banking & Finance, Volume 46, September 2014, Pages 43-58, ISSN 0378-4266, http://dx.doi.org/10.1016/j.jbankfin.2014.04.027.

[9] Tuckett, David, "Minding The Markets: An Emotional Finance View Of Financial Instability", Palgrave Macmillan", 2011.

[10] David M. Bholat, The future of central bank data, *Journal of Banking Regulation* (2013) 14, 185–194. doi:10.1057/jbr.2013.7; published online 3 July 2013; corrected online 14 August 2013.

Testing Big Data – do large data sets help reduce uncertainty?

David J. **Hand**, Imperial College London
Prateek **Buch,** Sense about Science
Phil **Bradburn**, National Audit Office

The advent of big data is often heralded as ushering in a new era of knowledge, bringing with it the ability to know more – and with greater certainty. But to what extent does big data help eliminate - or at least reduce - uncertainty in public policy choices? This panel discussion will set out how government policy and public discourse needs to acknowledge – and deal with – the uncertainty that is built-in to big data.

Sense About Science has previously brought together experts in diverse fields - climate science, clinical research, natural hazard prediction, public health and epidemiology – to help make sense of uncertainty. Drawing on their insights and bringing the discussion in to the big data era, this panel session will examine whether big, complex data sets help, or complicate, the task of using probability in public decision-making.

Speakers will address questions such as how we should interrogate big data in order to reduce uncertainty, whether big data techniques make it easier to plan for managing hazards, and whether big data makes it harder for the public to hold policymakers to account for how they use evidence.

As governments face up to significant challenges such as climate change and pandemic disease – and society considers how to regulate innovations that might help overcome these challenges – this discussion on managing uncertainty is vital if big data is to usefully inform the difficult choices ahead.

Using big data to map an innovative industry: the case of the UK video games industry

Juan **Mateos-Garcia** and Hasan **Bakhshi,** Nesta, UK

Big data sources provide the opportunity to reveal policy-relevant activities, behaviours and industries that were previously difficult to detect. In this paper, we illustrate this with a project where we used ‘big data’ in order to measure and map the UK video games industry.

Our main research question is thus:

Can a big data approach improve our ability to measure an emerging industry such as video games compared with traditional approaches based on SIC codes?

The video games industry provides an interesting case study of the challenges for measuring young, fast-growing, innovative industries:

In recent years, there has been a raft of policies put in place to support the UK video games industry, including educational reform and tax relief for the production of culturally British video games. There is however a general recognition that video games are inadequately captured in official government statistics, and this diminishes the effectiveness of these policies. Until 2007, the industry did not have its own Standard Industrial Classification (SIC) code, and it was therefore not possible at all to measure its size and evolution using official data. Although this situation has changed with the latest SIC code revision, there remain widespread concerns about company misclassification, linked to ‘reclassification inertia’ and usability issues in the business register.

Anecdotally, high levels of innovation in the sector bring with them additional measurement challenges. Innovative video games companies often work across industries and this creates uncertainty when choosing their SIC code. Digital innovation has resulted in the entry of large numbers of micro-businesses into the sector. This segment of the industry is however underrepresented in official business surveys, not least those that only sample businesses above a certain threshold size.

In the past, researchers have created data about the sector through surveys targeting samples drawn from company lists compiled by industry trade bodies and specialist consultants. Although this approach provides a highly relevant sample, and data, it is expensive to administer, and faces the risk of potential biases in its sample design.

In our video games mapping project, we set out to explore the potential of using online (big) data to identify UK video games companies more accurately, richly and cost-effectively than it has been done before. We did this by exploiting product directories, review sites and digital distribution platforms with detailed information about video games and the companies involved in their production. In theory, these data sources should help us identify video games companies on the basis of their economic activity, rather than the administrative information they supply when they become incorporated. We compare the outcome of this approach and of measuring the industry exclusively through ‘games’ SIC codes in order to address our research questions.

We scraped our data from seven product/industry online directories, resulting in a list of 226,302 unique game titles that led to the identification of 73,148 games companies globally. We fuzzy matched this company list against the API of Open Corporates (an open portal into international business registers) to identify businesses with a high probability of being UK based, resulting in a list of 8,880 matches that were subject to additional Quality Assurance. This stage included manual validation of a subset of our company list, a decision tree analysis to identify companies where manual validation was not undertaken, yet had a high probability of being in the games industry (on the basis of the data source where they were identified, their incorporation date, and their SIC code), and final QA by industry experts.. The final list included 1,902 games companies, for which we extracted additional data from the business register, including their SIC code and their location for mapping.

Our results confirm widespread concerns about misclassification (and therefore mis-measurement) of UK video games businesses: only a third of the companies in our dataset are correctly classified. Misclassification seems to be a more severe issue with younger companies, and companies operating in new platforms for games distribution. This is consistent with the idea that innovative businesses in particular are poorly captured by official video games SIC codes.

We also benchmark the results of using our ‘big dataset’ to map the video games industry across the UK with results based on business register data and official SIC codes. Although this comparison reveals a high degree of consistency in the places characterised by high agglomerations of games businesses according to both approaches, we find that official data misses a long tail of UK locations with some games-production activity – yet this information is important for national and local policymakers who want to understand the geographical distribution of the sector.

There are other important advantages in using a big data approach to measure the sector: it makes it possible to identify companies that are economically active yet too young to have selected a SIC code in the business register; it makes it possible to access data which is relevant for the sector but is not collected through official business surveys; and it provides a higher level of resolution (company level) than is possible with official sources (for reasons of data disclosure). We illustrate each of these advantages with findings from our research.

Using a ‘big data’ approach is not without challenges. We have had to develop new technological and analytical capabilities, not least to ensure that the outputs of the research are open and transparent (there is a growing ecosystem of big data analytics providers, but they generally use proprietary algorithms which limit the replicability of the findings and their usefulness for evidence-based policymaking). Quality assurance and communication of findings are also critical. The automated, fuzzy matching and classification of companies used in our research is probabilistic in nature, and a source of false positives/false negatives that are easy to spot in our highly transparent dataset. Manually removing such classification errors has high fixed costs, yet is important in order to maintain the policy credibility of the data set. We suggest that pragmatically combining automated data-collection processes and domain knowledge from industry experts can help big data researchers with policy audiences strike the right balance between scalability and reliability.

What the Internet of Things should learn from the biosciences

Boris **Adryan**, University of Cambridge

More than two decades have passed since a webcam monitored the use of a coffee maker in the Trojan Room in the Computer Laboratory at the University of Cambridge. Industry automation has for years used internet connections as replacement for physical wires in machine-to-machine (M2M) applications. The next big wave is a plethora of internet-connected devices that is coming into our homes: the consumer Internet-of-things (IoT) becomes reality.

While in M2M applications there is a distinct task and a defined business process that can be reflected in software, the consumer IoT is extremely vague: Which devices does a user own? What is the purpose of these devices? How should the information coming from sensors be used? What does the user want? In the extreme case, one man's temperature sensor could be another man's fire alarm... But how do you consider this in terms of data storage, provenance and organisation?

The biological sciences have seen a deluge of data and diverse data types for the past fifteen years. For example, prices are falling faster and yield of genome sequencing is rising faster than modelled by Moore's Law - a nice comparison to the 50M IoT devices predicted by some. Descriptions of life on the cellular and molecular levels themselves are inherently extremely complex. Researchers in the biosciences have agreed on data sharing standards, public repositories for raw and processed data, as well as common vocabularies and ontologies for the curation of such data (experiments, processes, biological concepts). This enables a freely available infrastructure that allows academic and industry users alike to 'mix and match' experimental data, infer knowledge on the basis of statistical analysis across different scales, and provide insight into biological systems that are bigger than just the sum of its parts.

In comparison to government spendings on HyperCat, a standard aiming to establish a 'catalogue of connected things', biomedical funding bodies have invested significantly more resource into enabling intelligent computational inference from biological data. For example, the most commonly used ontology to navigate molecular and cellular entities, Gene Ontology, has seen some $44M of investments since 2001 from the US National Institute of Health alone. Platforms for data sharing are largely paid for by governments, as an infrastructure freely available for both academia and industry.

Looking at the Internet as a primarily academic invention (with its later commercial success), the standards and platforms available for the biosciences (also accelerating industry research), it seems a worthwhile consideration to provide some core infrastructure for the IoT. Given the predicted importance and impact of the consumer IoT, we can and should not leave the development of infrastructure to commercial stakeholders alone. Big data policy makers need to recognise the analogies and governments need to make a concerted effort to create a digital environment that lets us leverage the opportunities of connect-able data!

Evidence based support to policy making in the Big Data Era: Understanding the context and learning from experiences in transitioning from ‘fat data’ to ‘big data’

Giuditta **De Prato**, EC JRC IPTS, Spain; Jean Paul **Simon**, JPS Multimedia, France; Jesus Vega **Villa**, EC JRC IPTS, Spain

The paper aims at marshalling facts about the notion of "big data", that have been spreading quickly without being, most of the times, properly defined thereby remaining all-encompassing. In doing this, the present work intends to put the phenomenon into a perspective, stressing the main challenges ahead: the economic, business and policy challenges. To do this the work would identify a series of examples of big data analysis application to techno-economic analysis aimed at addressing a number of policy issues with a socio-economic relevance, and a focus on the ICT industry in Europe, giving a measure of how much evidence based support to policy making can be changed by taking new data analysis into consideration. To introduce this set of case studies constituted by trial applications, an overview of the major initiatives taken recently by governments in the EU and the US in the approach to big data will be provided.

In order to give an overview of the drivers, and to take a look at big data’s likely trajectory, the paper starts by providing an introduction based on a review of the main sources available. Thus, the first section gauges the size of the data involved (market, volume), it offers glimpses of some of the assessment of the (potential) market linked to “Big Data”. Such section offers a tentative definition of what the somewhat fuzzy notion covers while tracing back an early definition identifying a three dimensional data growth (3Vs): volume (amount of data: petabytes or above), velocity (speed of data in and out needed for real-time collection/analysis of data), and variety (range of data types, formats and services, collected from a variety of collection mechanisms), to which more recently a fourth V for value was associated. It will introduce the basic components of “big data” as well as the value chain.

The second section questions the substance of the phenomenon trying to better underline its real present scope, to investigate beyond the present hype. It reveals that the benefits of big data are not always clear today (Forbes, 2013), that the amount of valuable useful data is still low. Besides it will highlight that despite industry hype, most organization have still to develop, implement or execute a big data strategy, as organizations continue to be wary of its impact. The section aims at identifying the barriers, tensions; power struggles and the time-span involved, explaining what the phenomenon really means, and fleshing out it real content through some field applications stemming from various industries. It also addresses the IT players, so-called “digital dragons” (firms like Google, eBay, LinkedIn, Amazon, and Facebook), firms that arguably were built around big data from the beginning. The paper considers other value added content providers then reviews some cases stemming from other industries, giving examples of applications to other sectors like banking, retail, telcos,…

The third section introduces some of the major initiatives taken recently by governments in the EU (2014) and the US (2012). It further deals with the threats the phenomenon may bring along but also with the opportunities in several areas. The

use of technology and data can both generate great value and create significant harm, sometimes simultaneously, therefore the section will explore how fully tapping that potential may hold much promise, as well as much risk. It also points at some of the technical hurdles that will have to be overcome.

Finally, the fourth section concentrates on the review of a "trial" project undertaken in 2014 in order to test tools and methods by applying them to real research issues so to test how a big data based approach could change the type and quality of the answers when techno-economic analysis is at stake. Each of the selected tools have been applied to trial micro-projects or tasks aiming at showing how evidence based support to policy making can switch to the 2.0 –or next – phase, thus encompassing dynamic access to data and advanced visualisation of “fat” data or “unstructured” data. Visualization techniques and data treatment used to improve the production of scientific evidence to support policy making such as in the case of the mapping of the European IT poles of excellence (formerly done in a previous research exercise by JRC IPTS IS Unit team for EC DG Connect) will be summed up.

The paper concludes delineating some potential policy interventions, and identifies the challenges ahead.

The work is based on desk research, a review of literature, review of the technical journals, analysis of annual reports, and meeting with experts and industry participants. Moreover, the proposed case studies encompass some relevant applications of new tools and methodologies carried out by the Information Society Unit and by the IT Department of the Institute for Prospective Technological Studies (IPTS) of the Joint Research Centre of the European Commission.

The views expressed are those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

Selected references:

Anderson, J., Rainie, L., (2012), The Future of Big Data, Pew Research, http://www.pewinternet.org/2012/07/20/the-future-of-big-data/

Atelier Paribas, (2013), Big data, big culture? The Growing Power of the Data and its Outlook for the Economy of Culture. Available at: http://www.forum-avignon.org/sites/default/files/editeur/EtudeATELIER_FA_2013.pdf

Bain, (2013), Big Data: The Organization Challenge. Available at: www.bain.com
Boston Consulting Group (2012),The Value of Our Digital Identity. Liberty Global Policy Series. Available at: https://www.bcgperspectives.com/content/articles/digital_economy_consumer_insight_value_of_our_digital_identity/

Brynjolfsson, E., Hitt, L.M., Kim, H.H., “Strength in numbers: How does data-driven decisionmaking affect firm performance?.” April 2011, available at SSRN (ssrn.com/abstract=1819486).

Cisco Visual Networking Index (2014), Global Mobile Data Traffic Forecast Update, 2013–2018. February 2014. Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

Davenport, T.H., Dyché, J., (2013), Big Data in Big Companies. International Institute for Analytics (iianalytics.com). Available at: http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

European Commission ((EC) (2014a), Towards a thriving data-driven economy. http://ec.europa.eu/digital-agenda/en/towards-thriving-data-driven-economy
e-skills UK/ SAS, (2013).Big Data Analytics: An assessment of demand for labour and skills, 2012-2017.

EMC Digital Universe study, (2014), http://www.emc.com/leadership/digital-universe/2014iview/index.htm
Ericson Mobility Report (2014). On the pulse of the networked society. www.ericsson.com/ericsson-mobility-report

Haire, A., J.,Mayer-Schönberger, V., (2014), Big Data - Opportunity or Threat, ITU GSR discussion paper, 2014.
IDC, (2012), Worldwide Big Data Technology and Services, 2012–2015 Forecast.
ITU, (2013). http://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx

Mayer-Schönberger, V., Cukier, K., (2013), A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt

Mac Vittie Lori, (2012). The Four V's of Big Data. Available at: https://devcentral.f5.com/articles/the-four-v-rsquos-of-big-data

McKinsey (2011), Big Data: The next frontier for innovation, competition and productivity.
http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx

SAP (2014), Beyond Connectivity. A Guidebook for Monetizing M2M in a Changing World,
SAS, (2013), 2013 Big Data Survey Research Brief. Available at: http://www.sas.com/resources/whitepaper/wp_58466.pdf

United States Executive Office, (2014a), Big Data: Seizing Opportunities, Preserving Values. Big Data Privacy Report,
http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

United States Executive Office, (2014b), Big Data and Privacy: A Technology Perspective.
http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

United Nations, UNPulse. http://www.unglobalpulse.org/about-new
World Economic Forum (WEF) (2013); Unlocking the Value of Personal Data: From Collection to Usage; Prepared in collaboration with The Boston Consulting.Group. http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf

Day 3 – Wednesday, June 17

Speaking up for us, the citizens: Whose policy, whose data, whose benefits?

Natasa **Milic-Frayling**, Microsoft Research Cambridge (MSRC)

Digital technologies are permeating all aspects of our lives. In contrast to other technological revolutions, such as printing and telecommunication that supported specific aspects of information sharing, digital technologies extend their reach and create new ways of data processing, knowledge acquisition, and information delivery. As such, they can create more powerful rifts in the society based on the accessibility of information and accessibility of computation.

This presentation adds to the voice of those who argue for establishing processes and practices to collect reliable information about the computing ecosystem and its stakeholders and disclose such information to the public. That is critical to enable informed social policies with citizens' engagement. Among key issues are the complexity of the computing systems and the designs that lack transparency of their workings. This makes reasoning about technology features and potential implications difficult for anyone but ICT professionals. Adding to that the intricate network of ICT providers and their intertwined business models, it is hard to determine where and how the value is generated without careful auditing mechanisms and expertise in business matters.

In order to illustrate these points we present empirical studies of information access on the Internet via Internet Browsers, facilitated through the standard and stateless HTTP protocol. The state management in the Browsers have been carefully devised by introducing browser cookies and providing guidelines which made provisions for user privacy. However, the practices that unfolded have given rise to real-time user tracking via third party cookies and a lucrative ad-bidding business that completely disempowers end-users. Most disconcerting is the imbalance between the computing innovation that enables real-time surveillance and bidding at unprecedented speeds and scales and a relatively low rate of user clicks and ad convergence. The latter indicates extremely low utility for the end user.

Perhaps more disconcerting are known vulnerabilities, such as the Browser cache design, that cannot be effectively addressed in any other way but through policy and law. Anyone can easily peak through the window in one's living room and take photos of inhabitants; yet, the society and the law make it clear that such behaviour is not acceptable. At the same time most of the citizens are not aware that anyone can break into their Browser cache and then pass on information about their browsing history to anyone else who may want to know about it.

Many data resources, big and small, have roots in the citizens' disempowerment, often through the lack of clarity about the nature of the service, change in the terms of engagement, lack of choice, or complete obscurity of the value exchange. Without a proper social voice to establish and defend citizens' rights, any government and society will be at the verge of technocratic dominance by the ICT industry and its businesses.

Governing by numbers

Richard **Harries**, Deputy Director, Reform

1. Government has always been done "by the numbers". Whether it was the thirteenth century Chancellor of the Exchequer counting jetons on his black and white chequered cloth or the 1941 survey of "Foundation Garments" commissioned by a War Ministry worried that demand for corsets was contributing to a nationwide steel shortage, the Government has always needed to understand the scale and dimensions of its remit.

2. More recently, particularly after the successful application of scientific techniques to the war effort during World War II, the Government has become increasingly interested in using numbers not just to understand the world around us but to influence it in socially desirable ways. Examples include:
    - economic forecasting and control of the money supply;
    - management of Next Steps agencies through key performance indicators;
    - formula-based grant distribution to hospitals, schools, councils and police;
    - Green Book policy appraisal, "Quality Adjusted Life Years" and PFI public sector comparators.

3. Yet despite this long history of measurement, analysis and prediction, the Government's record is less than impressive. Several weak spots appear to recur:

    - OLS regression models where the data are not normally distributed and/or where specific interpretations are applied to the residual terms. (The latter typically occurs in league table analysis and/or to determine the technical/ allocative efficiency of decision making units.[3])
    - Failure to account properly for inherent uncertainty through the use of sensitivity analysis and/or the critical dependence of models on challengeable assumptions.[4]
    - Time series projections/predictions where the generating models rely on underlying continuity in the data series – even though it is often the very structural breaks in these data sets that are of most interest to policy makers. (For example, the impact of 2008 credit crunch on Bank of England inflation forecasts.)
    - The use of datasets, originally collected for one purpose, for wholly different purposes – typically where administrative datasets are used for performance management and/or incentives-based regimes. This

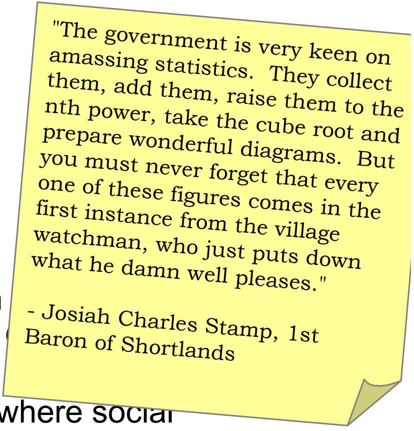
[3] Even where relatively sophisticated statistical models are applied, such as Stochastic Frontier Analysis, it is necessary to make untestable assumptions about how to partition the residual term into an error function and a frontier function.

[4] An example of the latter is the choice of discount rates in the Stern Review on the Economics of Climate Change.

approach frequently reveals more about weaknesses in the dataset than it does genuine differences in performance.[5]

4. This talk will explore these issues and ask:

    - are there new approaches to data collection, manipulation and analysis that offer technical improvements to the way deals with numbers?
    - where technical improvements cannot be made, perhaps where social problems are inherently unmeasurable, what are the appropriate structural responses?

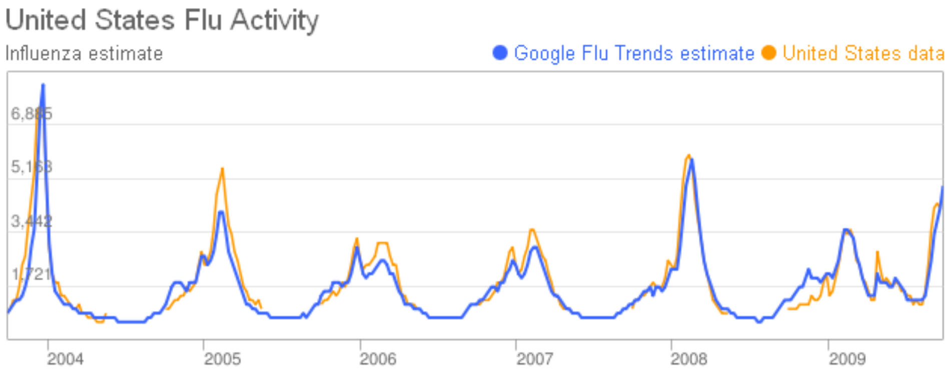


"The government is very keen on amassing statistics. They collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn well pleases."

- Josiah Charles Stamp, 1st Baron of Shortlands

Technical improvements

5. The sorts of questions these issues pose include whether new tools and techniques are available to improve the situation and/or whether new technology offers wholly different approaches to tackling well-worn public policy questions. Recently, for example, there have been intriguing suggestions that the use of "big data" can transform the way companies capture and analyse huge volumes of data to allow radical customisation, constant experimentation and novel business models. A well known – but now discredited – example of this latter approach is Google Flu Trends, which was based on the simple insight that there appeared to be a close relationship between how many people search online for flu-related topics and how many people actually acquire flu-like symptoms.



United States: Influenza-like illness (ILI) data provided publicly by the U.S. Centers for Disease Control.

Structural responses

6. Where measurement and analysis cannot be improved upon, the next question must be what the appropriate response of Government should be. It could be to continue as present, muddling through with incomplete and often inconsistent knowledge. However it might also lead to questions about the shape and structure of the governmental process itself. What structural

[5] It can also lead to undesirable gaming behaviour, for example by waiting until the end of the sampling time frame to report progress, thereby artificially inflating the subsequent growth rate.

solutions exist, consistent with modern democratic society, that are able to deal with this sort of inherent uncertainty?

7. Potential solutions include:

    - transfers of power/responsibility to different spatial levels of government (for example, through decentralisation)
    - use of the price mechanism to reveal unstated preferences (for example, through privatisation)
    - multiple redundancy mechanisms to maintain stability at critical turning points (for example, through a system of checks and balances)

Whitehall Monitor: charting government in 2015

Gavin **Freeguard**, Senior Researcher working on Whitehall Monitor, Institute for Government

The UK regularly tops global open data indices and is rightly regarded as a leader in the field. But data – a series of numbers in spreadsheets – is one thing; information – which expresses it in an understandable form – is quite another. The Institute's *Whitehall Monitor* regularly takes government data and uses it to understand the size, shape and performance of government. It also highlights where government could improve the transparency and accessibility of information, including the impact it makes on the real world.

How do UK voters want to engage with Parliament?

Finbarr **Livesey**, University of Cambridge, UK

There has been much talk on the digitization of Parliament, especially following the Digital Democracy Commission's recent report. However, it is unclear that providing digital pathways to engagement will encourage the public to be part of the policy making process. The promise of strong open policy making may be addressing a need that doesn't exist.

This paper reports the results and analysis of a large scale survey of UK voters (N=1,676) which investigates the levels to which the public feel they are involved and how much they wish to be involved in policy making in the UK Parliament. It also reports on the channels that are currently used and which channels (such as Twitter or face to face meetings) are preferred. Finally the paper highlights where there are significant differences by voting intention, age, gender and social grade for each of these questions, highlighting that there is no one size fits all digital approach and that any approach in isolation is going to disincentivise some parts of the electorate.

See what a state the maintenance of "state" has got the State in!

John **Taysom**, RVC, UK

The internet and the Web are now essential tools for policy delivery. Analysis of internet and Web activity also have growing relevance across the policy spectrum from medicine to security. But as the State moves to incorporate internet and Web into policy decision making and delivery in fields like education and medicine the issue of personal privacy will continue to recur as a theme; often as an inhibitor to progress, and often portrayed as an 'either or decision'. But what if we could have both? Better policy decisions and cheaper more effective policy delivery and yet retain personal privacy? A new approach, now the subject of EU and US granted patents, makes use of the observation that what is valuable often in policy making (and for that matter valuable in commerce) is not the specifics of who you are but in what relevant groups or crowds are you clustered, at a point in time, within a given context, at a certain location. By applying well understood heuristic clustering techniques, but with a lower limit on the cluster size, it is possible to ensure that the crowds are "big enough" (in a formal, measurable, auditable sense)to make identification of the individual difficult (in a formal sense). This means the data utility in personal data need not be at odds with privacy. This is a novel approach with potentially exciting applications in Public Policy.

The work to arrive at this conclusion and to begin to fit the solution, designed originally for commerce, to questions of public policy was done on a two year cross-school Fellowship at Harvard University and is being continued at CSaP as a 2015 Policy Fellow.

The first practical applications of this approach, dubbed "3 is a crowd" are being developed by Cambridge-based Repositive Ltd. (for safe sharing of genomic data) and London-based Privitar Ltd. for 'corporate Big Data' in particular data from connected cars and IoT more generally. Other applications include smart energy metering, transport monitoring, analysis of cell-phone traffic, and others where personal privacy and data utility could otherwise be seen as antagonistic.

Policy considerations for Responsibility-in-Use: Mediating Big Data in the Public Sphere

Roxane **Farmanfarmaian**, University of Cambridge-al-Jazeera Media Project, POLIS, University of Cambridge

As Big Data is increasingly being employed not only by journalists, but by computers able to sidestep journalists in writing and communicating news, the question arises whether information is superseding knowledge repertoires as a driver and definer of news generation. Further, where the former has in the past been understood to be more neutral than the latter, is that still the case, when information is extracted through the process of Big Data generation? How vast numbers are collected and why are critical to understanding the material being presented in the public square, as are the contexts in which interpretations are situated. Thus, Big Data, as a conglomerate of numbers, cannot be understood as neutral but instead, able to carry meaning through the structure and purpose of its generation. Presentation of Big Data as ‘fact’ de-narrativizes the process, with important implications for interpretation, and presentation. What is more, two elements within the process of Big Data collection contribute to homogenization: first, the preservation of privacy and the need to extract from the ‘unseen’; second, the value of the trendline, which creates safe spaces for normative conformities, but challenges and erases the existence of alternative, minority or ‘resistance’ identities or behaviours. Bruno Latour (1988) has argued that science is often politics by other means. Those actors who control the knowledge produced by science, but also the legitimising language of science, can thereby gain power over their adversaries.

Michel Foucault (1977) advances the idea that certain techniques and institutions have converged within the context of the modern state to create systems of disciplinary power. The Internet has a panoptical quality in that users can rarely be sure whether their online activities are being monitored or their activities are being hoovered up into collections of data. Moreover, what tools of resistance are available to the user require high levels of technical knowledge. Whilst the Internet is a mass-public medium, its technological basis presents technical and conceptual complications that present a paradigm shift, in that they far exceed previous means of access to information. A troubling contradiction regarding users’ relationship with the Internet thus emerges. Following Alexis Artuad de Ferriere (2014), the operation of the Internet depends upon open and transparent protocols, many of which are, technically, verifiable and even modifiable by the public at large. However, the reality is that the vast majority of users neither have the knowledge nor the inclination to verify the terms of their Internet usage, or to monitor how others might utilize information about their usage. This contradiction exemplifies an eroded Enlightenment myth: that the development of science and technology occupies a distinct space from that of politics (Shapin and Shaffer, 1985). As noted by Lyon (2011), the information society is also the surveillance society, and science is often politics by other means.

This has significant implications for the use and interpretation of Big Data in journalistic practices, where sourcing, filtering, and presentation by technical experts are given context and interpretation by nontechnical experts to construct meaning in the mediation of knowledge to the public. This is of particular concern for if Castells is correct, transformation of people’s minds is the core source of influence in today’s world, and the media, as a distributive network of images, reportage and narrative, is the most effective in doing so. This study, therefore, interrogates Big Data use as a resource that is shared with few normative restraints and yet occupies a special value by virtue of size, source and/or gestation speed, in how it is mediated and how it is received by the target audience.

This suggests that the employment of Big Data in mediated knowledge and meaning generation requires responsibility-in-use, based on normative considerations of constraint backed up by regulation and institutional remits (legal regimes that determine how much access the state, or large corporations, have to users’ data). This study investigates how Big Data appears in news and analysis stories in the press (both conventional and online), and by employing a critical discourse approach, highlights the contested nature of the information it represents, as well as its affect on the narratives it resources within mediated and contextualized journalism. The paper then offers suggestions for policy approaches that address the issue of responsibility-in-use. These draw on precedents of internet governance (laws that exist to uphold freedom online), as well as journalistic norms of good practice.

The Representational and Ethical Limitations of Using Social Media Data Real-Time for Policy-Making

Ella **McPherson** and Anne **Alexander**, University of Cambridge

This paper is based on written evidence we submitted to the UK Parliament's Science and Technology Committee for their 2014 'Social Media and Real Time Analytics Inquiry.' We were concerned, in particular, about the call for evidence's implication that social media data might be used in governmental decision-making through real-time analytics and as a substitute for other, researcher-generated data. We wished to urge extreme caution, drawing on our own research into social media use in contexts of conflict, and based on the representational and ethics limitations of using social media data in this way.

First, we argue that social media data is not directly representative of facts offline because it is subject to a variety of distortions. This includes those arising from the commercial nature of social media platforms; the fragmented nature of social media data and the difficulty in definitively establishing the source, place, and time of production as a result of the lack of cues and context for the reader; and the emergence of social media platforms as important sites where social and political conflicts play out.

Second, we argue that, in order to use social media data to establish a fact, it must undergo a verification process. This verification process requires time, triangulation (cross-referencing using a variety of sources and methods) and human expertise. Verification is thus incompatible with real-time analytics of social media data. Furthermore, because verifying social media data may involve identifying its source, this approach raises ethical concerns related to accessing data, informed consent, and anonymization.

We suggest that an alternative to social media in this way is to consider it instead as an 'awareness system' (Hermida, 2009), which could indicate areas of interest for directing further research and investigations. However, such an approach must, we believe, minimize the collection of personal data and requires the triangulation of research methods and sources before any actionable claims are made as to the representativeness of social media data.

The paper concludes by considering developments in the use of social media by the UK government in the intervening time.

Narrating Networks of Power: Narrative Structures of Network Analysis for Journalism

Liliana **Bounegru**, Jonathan **Gray** and Tommaso **Venturini**

In an era of Big Data, networks have become the core diagram of our age. As popular books on the topic contend, the concept of networks has become central to many fields of human inquiry and is said to revolutionise everything from medicine to markets to military intelligence.

In the context of media and journalism, using data to map networks is praised for its potential to expose the workings of power, be it financial or political. The work of the artist Mark Lombardi, as well as power mapping projects such as They Rule, Muckety, Little Sis, Poderopedia and the Organized Crime and Corruption Reporting Project's Visual Investigative Scenarios have opened up journalistic imagination about how network analysis and mapping might be used in the service of journalism.

While journalists have been experimenting with network analysis and mapping to discover and tell stories with data for decades, the breakthrough moment of this analytical and storytelling device in journalism has yet to come.

Journalists have been reluctant to embrace network analysis and visualisation, and not without good reason. While network analysis can be an effective exploratory tool, in order to be used as narrative tools networks have to be embedded in a rich conceptual framework to generate meaning.

In this article, we propose a possible framework to breathe meaning into networks, a vocabulary of narrative functions that network analysis can play, based on the popular social research approach of 'issue mapping', and on examples of use of network analysis and mapping techniques in journalism. Developed at the crossroads between Science and Technology Studies and Internet Studies, issue mapping operationalizes concepts from Actor-Network Theory (ANT) in order to study the state of public issues.

The resulting classification of narrative structures of network analysis in journalism and issue mapping will provide an opportunity to reflect on the potential and limitations of network analysis for mapping power in the context of journalism, as well as on how essential aspects of journalistic epistemology – such as notions of time, space and narrative – are being reconfigured by this set of technologies, practises and concepts.

Are we measuring the right things? From disclosure and data portals to participatory data infrastructures

Jonathan **Gray**, Open Knowledge Foundation

Over the past few years open data has transitioned from being a niche area in legal, policy and tech circles to being a prominent topic on the global political stage. At the heart of the "open data revolution" are open data portals – from the initial releases of data.gov and data.gov.uk to hundreds of local, regional, national and international data portals around the world. Quite a few of these data portals aspire to gather feedback from their users on priorities for digital data release, and some have started to experiment with mechanisms for enabling users to improve or contribute data. However, critics of open data contend that there is a danger that open government data programmes focusing on proactive dislosure can correspond with a weakening and de-emphasis of access to information regimes as an instrument to providing citizens, civil society and other actors with the information they need (Janssen, 2012).

This paper will present ongoing research on how open data intiatives, as complex socio-technical assemblages, engage with users around what data should be published and publication priorities – including an examination of policies, practises, technologies, software and strategies, drawing on interviews with practitioners and advocates, and the analysis of key documents and digital artefacts. It will look at the role of notions of “transparency” and “disclosure” in open data initiatives, and argue for a broader conception of “participatory data infrastructures” which deploy a range of different practises, instruments and technologies in the service of pro-actively engaging with data users not only about what information should be released, but what information should be collected and generated in the first place. It will draw on recent advocacy work around the “beneficial ownership” of companies and tax base erosion, as well as recent research on “data activism” and “statactivism” (Isabelle, Emmanuel and Tommaso, 2014).

References

Isabelle, B., Emmanuel, D. & Tommaso, V. (2014), "Statactivism: forms of action between disclosure and affirmation", in Partecipazione e conflitto. The Open Journal of Sociopolitical Studies, vol. 7, n. 2, pp. 198-220.

Janssen, K. (2012) “Open Government Data and the Right to Information: Opportunities and Obstacles”. The Journal of Community Informatics, Vol 8, No 2.

From Big Data Sets to Collective Human Behavior Patterns and Urban Spatial Structure - Analyzing and Simulating Spatial-Temporal Dynamics in Shanghai

Chaowei **Xiao** and Elisabete **Silva**, University of Cambridge, UK

Using traditional approaches, such as surveys, GPS navigation, to explorer urban spatial structure and transportation flows through a city are not only time consuming, in-accurate, expensive, but also have the limitation on smaller samples. Only in recent years, the research on ‘big data’ in urban geography and planning fields has been introduced; diverse new sources of big data appeared during the last years, such as mobile-phone data and smart card data.

For instance, according to the signal data sent by mobile-phone and the base station location, the approximately spatial position of mobile-phone user can be calculated. These kinds of data have a lot of advantages, such as more efficiency during data collection/processing, real-time, and have better spatial resolution. Due to the current high rate of mobile phone holders, in Shanghai, for example, at the end of 2012, the number of mobile phone users in Shanghai reached 30.083 million with the total population of 23.80 million, as a consequence of this vast amount of information, the human spatial behavioral patterns and urban spatial structure can be easily obtained.

This research presents an approach to explore urban spatial structure and human behavior patterns using mobile phone signal (positioning) data in the case study of Shanghai.

Existing studies demonstrate that big data such as mobile-phone data and smart card data can be used to identify the urban spatial structure, can reveal the spatial-temporal dynamics of transportation flows and can be useful for urban planning and management. Ratti et al (2006) use mobile-phone data of Milan, Italy, to describe the spatial distribution of mobile phone users in different time periods, the spatial distribution of activities and intensity in different time periods. Morency et al. (2007) successfully measure spatio-temporal variability of transit trips in Gatineau, Canada, based on big data in that city. Krisp (2011) based on the spatial distribution's density of mobile phone users in the night, estimates the spatial distribution of the resident population and uses these results optimize fire station and other emergency facilities planning. Becker et al (2011) based on the data of mobile phone users in different time of the day, analyzes the commuter population and employment space in New York sub-urban area. Ying Long et al. (2013) use the smart card of Beijing to characterize spatial and temporal mobility patterns of the city. Batty et al (2013) collected and analyze the data of the London underground system (Oyster card data), enabling people to infer the statistical properties of individual movement patterns in a large urban setting. Thomas Holleczek et al (2014) using a data mining approach based on 24 hours phone-based data of Singapore, visualized the mobility and connectivity of Singapore. In general, the mobile-phone data analysis for urban geography and planning registers increasing research, nevertheless although some progress, we are still in early ages of processing and evaluation of results.

This paper/presentation will focus on mobile-phone data for Shanghai. Shanghai is located in the southeast coast of China and the largest city in China. Shanghai is a global financial center and a transport hub with the world's busiest container port. It is one of the most compact city in the world, and the land resources are extremely scarce in Shanghai. So, the analysis of population distribution and resource allocation is important in Shanghai. Meanwhile, the Shanghai master planning 2040 is in progress, this research focuses on Shanghai and the input data for our study was a prerequisite data set collected from Shanghai Master planning(this data set includes: Shanghai mobile-phone, Signal Data, Road network database, Underground Railroad, population census and economic census in Shanghai).

The first part of the research presents an approach to explore urban spatial structure using mobile phone positioning data in Shanghai. The initial methodology uses the location data of mobile phone and base station, and applies data mining techniques to evaluate the number of users connecting to each base station, summing the number of users and producing density maps of the mobile phone users - generated by kernel density analysis or other geo-statistic methods. The urban spatial structure can also be measured by dynamic matrices.

The second part focus on identify living and working place of Shanghai. We can generate multi-day average user density maps from workday 10:00 to 23:00 on work days and from 15:00 to 23:00 on weekends; once this process is accomplished, then, by spatial cluster and density classification of density maps, the ranks and functions of “hot spots" in Shanghai can be identified. Lastly, we can identify residential areas, employment areas and leisure areas in Shanghai and measure the respective level of the three functions in the “hot spots” by comparing the ratio map of day and night user density.

In a third part, using this big data, human behavior patterns can be generalized. We can use each individual mobile phone data set to generate the real O-D distribution of Shanghai. Then, we can derive some patterns by using geo-spatial methods. After that, we can derive some patterns by using geospatial methods from the individual data. Patterns can be identified such as A) Intercity long distance transportation working pattern. B) Cross-district long distance transportation working pattern. C) Short distance transportation working pattern. D) Job-housing balance pattern. E) No working pattern. F) Random working pattern. G) Two or more living place pattern.

Finally, in further stages of the research, the relationship between the spatial population flow pattern and land use\road\underground\planning policy\space syntax can be explored. For instance, to incorporate human behavior patterns with ABM model to simulate the future population spatial distribution and future O-D distribution in Shanghai. The human behavior patterns, the future human spatial distribution and future O-D distribution also can be simulated.

The findings of this research are expected to provide support for policy formulation to alleviate traffic pressure, reduce carbon emission from transportation and optimize urban spatial structure of Shanghai.

Using mobile-network big data for urban & transportation planning in Colombo, Sri Lanka

Rohan **Samarajiva** and Sriganesh **Lokanathan**, LIRNEasia, Sri Lanka

As the world becomes more urban, planning and efficient operation of urban infrastructure and services become more important. Until now, planners and managers relied on periodic and expensive surveys, if they used data. Alternatives are emerging because of changes in data storage and processing, known by the shorthand of “big data.” In a “smart city,” information-communication technologies (ICTs) are used to enhance feedback loops within the complex system that constitutes the city.

The ubiquitous mobile phones and citizens themselves can serve as the primary sensors. As a person with a mobile phone moves through the city, the mobile phone communicates with Base Transceiver Stations (BTS) so the network can complete a call if one comes through. These records are called Visitor Location Register (VLR) data. They are massive in volume and are usually written over. Smaller in volume are Call Detail Records (CDR), which are generated every time a call is made and received, and SMS is sent, the Internet is accessed and prepaid value is loaded. LIRNEasia has obtained historical, anonymized CDRs from multiple mobile operators in Sri Lanka in order to conduct exploratory research on the insights that may be gained for urban and transportation planning.

Findings on the following topics will be presented:

- Different kinds of land use can be identified from the diurnal loading patterns of BTS, enabling closer alignment of plans and actual land use.
- Changes in population densities relative to midnight ("home location") may be mapped for different times. Daytime or "work" locations may be identified, along with where people came from.
- Communities, defined in terms of interactions within the community being greater than those outside, can be identified. Some of these "real" communities overlap with administrative boundaries; many do not.

Discussion of the potential benefits of mobile network big data must be balanced by consideration of potential harms. Building on long engagement with utility customer data issues, LIRNEasia has prepared a set of draft guidelines that seek to lower the identified barriers to release of mobile network big data to third parties for research of the type discussed above.

Big data and the monitoring of post-disaster economies

Timothy **Wilson**, United Nations Economic Commission in Africa (UNECA), Rwanda

I am developing a monitoring framework for post-disaster economies. My presentation will explore the opportunities and challenges that the Big Data revolution presents for such monitoring.

THE FRAMEWORK

Disasters are common. There were 337 disasters caused by natural hazards in 2013, according to the IFRC. In terms of disasters caused by man-made hazards, the Syrian conflict alone has created nearly 4 million refugees according to UNHCR.

Successful economic recovery from such disasters requires consistent, coherent monitoring. This enables informed decision-making and provides a way of assessing progress. However, there is no commonly used framework for such monitoring. If it occurs at all, economic monitoring tends to be designed and implemented on an ad hoc basis. Lessons from theory, empirical research and other disasters are often missed.

This occurs despite the fact that economic stability plays an important role in disaster recovery. In post-conflict environments, for instance, (Miguel, Satyanath, & Sergenti, 2004) demonstrated that improved economic prosperity, all else equal, reduces the propensity for a reversion to conflict. Widely-used humanitarian monitoring frameworks, such as those produced by OHCA, have shown that post-disaster monitoring frameworks are possible, even in challenging circumstances. Furthermore, innovations in the collection, processing and management of large amounts of electronic data (Big Data) are producing new opportunities for measuring economic variables.

I am developing a framework for monitoring economic recovery in post-disaster environments. The framework lists variables to be monitored. It also explains how the data can be collected and analysed for each variable. The framework and its methodology are based on theory, case studies and empirical research. The objective is to inform policy and decision making, as well as provide early warning signals and serve as a mechanism for contextualising progress. A unified framework would minimise duplication while also providing a structure for consolidating knowledge on post-disaster economic recoveries.

MY PRESENTATION

My presentation introduces the draft framework and then focuses on the opportunities and challenges presented by Big Data.

With respect to opportunities, I describe a number of ways that Big Data has been used in post-disaster economic recovery. For instance, it was used to monitor shifting populations and centres of commerce in New Zealand, following the Canterbury earthquakes of 2010 and 2011. It provided a way of measuring transport usage in West Africa. I have also used large panel datasets to gain a window into otherwise un-measured economies, such as Somalia. I discuss the importance and limitations of each of these opportunities.

With respect to the challenges, I describe the difficulty of maintaining confidentiality, which was a particular challenge in post-earthquake New Zealand. I will also explain the importance of ensuring that large amounts of data are accompanied by large amounts of analysis. Data without analysis can be costly. I will conclude with a brief discussion about the relationship between official statistics agencies and Big Data, particularly in the context of low-income countries recovering from major disasters.

Natural History Museums in Europe as hybridized research infrastructures: faded glory or a digital phoenix rising from its ashes?

Hank **Koerten** and Peter van den **Besselaar**, VU University, Netherlands

1.Introduction

Biodiversity has become a hot societal issue; almost every right-minded person has worries about global pollution and its impact on nature. These notions have been put to action, both politically and scientifically, on national and global scale (Brown, 1994; Gaston & Spicer, 2004; Watson et al., 1995). While the problem of diminishing biodiversity is often perceived as being clear and present, it has obvious notions about the past, due to the mere fact that we compare the current situation with the past, for which historical information is needed (May & Beverton, 1990).

Usually we think of information residing in publications as these sources can provide us with clues about changes in biodiversity. To make more accurate evaluations, however, other sources of information are needed, which has been articulated before:

'Unfortunately, the historical acceptable products of research - namely, peer-reviewed publications - typically do not contain the unreduced data or information

necessary for new analyses by future scientists.' (p. 310)(Ingersoll, Seastedt, & Hartman, 1997).

This quote addresses the problem that every biodiversity researcher is facing: where to find appropriate data to identify changes in biodiversity.
At the same time, natural history museums in western countries have vast collections of specimens with information on species spanning a century or more (Winker, 2004). They have a rich and vast history of standardized specimen collections that has been the source for scientific research on biodiversity for years (Allen, 1976; Barber, 1980). To utilize natural history collections would seem logical, assuming they continue to take the role they have been playing for centuries.

However, today we see two developments affecting the future of natural history museums in biodiversity research. First, natural history museums increasingly have become insecure about their future (Lane, 1996; Lister, 2011; Ponder, Carter, Flemons, & Chapman, 2001; Shaffer, Fisher, & Davidson, 1998; Winker, 2004). The objectives of natural history collections are still clear: informing the public and being a source for scientific research. However, the discussion reveals concern about diminishing time and effort devoted to scientific research within natural history museums. Second, curators and scientists in natural history museums tend to engage in practices going beyond the scope of their own institution (Duin, King, & Van den Besselaar, 2012; Duin & Van Den Besselaar, 2011; Smith & Duin, 2009; Van den Besselaar, Koerten, & King, 2013). Recent EU-funded programs like EDIT, ViBRANT and Synthesys were meant to stimulate cooperation between biodiversity researchers, with an agenda-setting and coordinating role for jointly operating natural history museums in Europe establishing novel digital infrastructures.

Both developments are triggers for change of strategic orientation in natural history museums, which might affect collection management, but might also affect visitor relations. How are these developments experienced at the shop floor of natural history museums? How do curators and scientists perceive the future of their collection being part of their natural history museums? Do they engage in external cooperation? Does this affect their relationships with their national audience? Such questions can be boiled down to a more conceptual, encompassing research question: does the taxonomic workforce of natural history museums consider their collections as part of a scientific infrastructure? This question guides our ethnographic research on curators and scientists in natural history museums. We want to understand the relationship between natural history collection and the wider context of biodiversity research.

2. The natural history collection as a research facility

People have always loved to go out in the fields to enjoy nature. They expressed their appreciation by talking and writing about it or by bringing home plants and dead animals as trophies (Allen, 1976). In the 19th century, people actively involved in nature became called naturalists, interested in their own neighborhood (Barber, 1980). So-called field-naturalists were collecting specimen by treating them as hunting trophies, closet-naturalists were focused at managing a collection: a herbarium or multiple stuffed, prepared animals. (Allen, 1976). Natural history provided the paradigm to study nature in Western civilization, being immersed with

Christian values, based upon biblical stories (Barber, 1980). This orientation explains the initial halfhearted reception of Darwin’s insights of natural selection.

The majority of natural history museums as we know them today have been established in the 19th century, based upon above described collections, withdrawn from ‘cabinets of rarities’ with sculptures, paintings, furniture, botanical gardens, stuffed animals and other specimens, previously owned by wealthy citizens (Allen, 1976). Soon it became clear these collections were unstructured and poorly managed, neither of interest to scholars nor the general public. Specialization became the name of the game: dedicated natural history museums, being organized at a national scale were subdivided into departments specializing along the lines of specimen classification: botany, entomology, ornithology, zoology, etc. With constantly expanding collections it became important to have the standardized, following scientific rules. It also became clear that it was impossible to have the entire collection permanently on display. The scientifically sound collection became the base both for changing exhibitions serving the general public and for research (Griesemer, 1990). Slowly but steady, natural history museums became seen as having a collection representing changes in nature which was suitable for the study of evolution (Griesemer, 1990).

From 1880 onwards, the idea of a natural history museum being beneficial to society became manifest. A specimen collection was seen as contributing to health, education and economy, and when studied by scientists, knowledge was produced. Studies by curators and other scientific staff used to be published in guides and catalogues in order to present sub-collections into a meaningful whole (Star & Griesemer, 1989). Thus the insights of Darwin were tested, extended and specified through systematic collection and ordering of specimens.

Collections could also play a role in demonstrating the unity and/or sovereignty of a specific state: the Berkeley’s Museum of Vertebrate Zoology held a collection signifying the importance of the western part of the US and the state of California in particular (Griesemer, 1990), the Natural History museum in London held a collection attempting to cover the British Commonwealth, while the Royal Botanical Gardens at Kew had a collection aimed at ‘ economic botany’ supporting colonial trade of staple products (Brockway, 1979).

Today, natural history museums have developed themselves as multi-purpose organizations where brains, databases, funds and facilities are organized to be hotspots of biodiversity research. Still based on traditional skills and accommodating facilities, they try to adapt to the requirements of the digital age (Smith & Penev, 2011).

References

Allen, D. (1976). The naturalist in Britain: a social history. London UK: Allen Lane/Penguin Books.

Barber, L. (1980). The Heyday of Natural History, 1820-1870. Garden City NY: Doubleday

Brockway, L. (1979). Science and colonial expansion: the role of the British Royal Botanic Gardens (Vol. 6). New York NY: Academic Press.

Brown, J. (1994). Grand challenges in scaling up environmental research. In W. Michener, J. Brunt & S. Stafford (Eds.), Environmental Information Management and Analysis: Ecosystem to Global Scales (pp. 21-26). London UK: Taylor & Francis.

Duin, D., King, D., & Van den Besselaar, P. (2012). Identifying Audiences of E-Infrastructures-Tools for Measuring Impact. PloS one, 7(12), e50943.

Duin, D., & Van Den Besselaar, P. (2011). Studying the effects of virtual biodiversity research infrastructures. ZooKeys(150), 193-210.

Gaston, K., & Spicer, J. (2004). Biodiversity: an introduction; Second Edition. Malden MA USA: Blackwell Publishing.

Griesemer, J. (1990). Modeling in the museum: On the role of remnant models in the work of Joseph Grinnell. Biology and Philosophy, 5(1), 3-36.

Ingersoll, R., Seastedt, T., & Hartman, M. (1997). A model information management system for ecological research. BioScience, 310-316.

Lane, M. (1996). Roles of natural history collections. Annals of the Missouri Botanical Garden, 83, 536-545.

Lister, A. (2011). Natural history collections as sources of long-term datasets. Trends in ecology & evolution, 26(4), 153-154.

May, R., & Beverton, R. (1990). How many species? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 330(1257), 293-304.

Ponder, W., Carter, G., Flemons, P., & Chapman, R. (2001). Evaluation of museum collection data for use in biodiversity assessment. Conservation biology, 15(3), 648-657.

Shaffer, H., Fisher, R., & Davidson, C. (1998). The role of natural history collections in documenting species declines. Trends in ecology & evolution, 13(1), 27-30.

Smith, V., & Duin, D. (2009). Scratchpad survey 2009. London UK/Paris F: Natural History Museum/Muséum national d'Histoire naturelle.

Smith, V., & Penev, L. (2011). Collaborative electronic infrastructures to accelerate taxonomic research. ZooKeys(150), 1.

Star, S., & Griesemer, J. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkely's Museum of Vertebrate Zoology, 1907-39. Social Studies of Science, 19, 387-420.

Van den Besselaar, P., Koerten, H., & King, D. (2013). Suggestions for potential user groups and audiences for Scratchpads Vibrant (pp. 47). Amsterdam: VU University.

Watson, R., Heywood, V., Baste, I., Dias, B., Gamez, R., Janetos, T., . . . Ruark, R. (Eds.). (1995). Global biodiversity assessment. Cambridge UK: Cambridge University Press.

Winker, K. (2004). Natural history museums in a postbiodiversity era. BioScience, 54(5), 455-459.

Predicting sense of community and participation by applying machine learning to open government data

Alessandro **Piscopo**, University of Amsterdam, Centrum Wiskunde, Informatica; Ronald **Siebes**, VU University Amsterdam; and Lynda **Hardman,** Centrum Wiskunde, Informatica, Netherlands

Community capacity (CC) is the ability of people to undertake, collectively or individually, any action that benefits their community [3]. Since a higher CC gives a greater indication of success for community programmes [3], it is useful to measure it for policy makers and local admnistrators. However, such measurement is often too onerous, as it is predominantly performed through locally organised surveys [3].

We studied an approach allowing reliable and inexpensive measurements of CC dimensions. We investigated to what extent these could be predicted by applying a machine learning algorithm, called Random Forests, to open government data not describing social dimensions, or secondary data.

We applied this technique to two dimensions: sense of community and participation. In both cases, the predictive models created yielded a high accuracy (R2 was respectively 76.5% and 62.5%). The variables contributing the most to prediction accuracy were only partially in agreement with the most influential factors on sense of community and participation found in the scientific literature.

RESEARCH QUESTION
Our main research question is: to what extent can we produce measures of sense of community and participation by applying machine learning to secondary open data? Subsidiary questions are which variables contribute most to the measures and whether they agree with those stated in the literature. Two criteria were set for our measures: consistent nationwide applicability, i.e. availability for any area within our context (England); high geographic precision, i.e. neighbourhood level detail.

RESEARCH METHOD
We used machine learning because it is not yet in wide use in the area of our study, notwithstanding its promising results in several fields. Second, machine learning models rely on discovering patterns inherent to the data [1], being more easily adaptable to different contexts. Random Forests was chosen due to its characteristics: ability to generalise beyond the training set; suitability for small datasets; high prediction accuracy; ability to provide a measure of the contribution of each variable to the prediction accuracy [2].

For each model, we selected a number of potentially relevant independent variables and a dependent variable, i.e. an existing measure of sense of community or participation, to train the model. The geographical coverage and level of detail of the dependent variables had to be as close as possible to the ones desired for the predicted measures.
The independent variables were selected by first identifying relevant indicators for sense of community and participation in the literature, e.g. socioeconomic conditions or ethnicity. Second, we mapped each indicator to variables within the sources providing data with the same geographic coverage and detail and corresponding time range as the dependent variables.

DATA COLLECTED
The dependent variables were the National Indicators from 2008 Place Survey NI 002 for sense of community and NI 003 for participation. Their geographical coverage is the whole of England, with local authority as the most detailed level available. The two models created included data spanning a time range from 2008 to 2011, with same geographical coverage and detail as the dependent variables, from a total of 23 datasets collected (17 from 2011 Census).

DATA PROCESSING
To answer our main research question, we trained our models using the Random Forests algorithm and evaluated their accuracy on the basis of the root-mean-square error (RMSE), a measure of the difference between predicted and observed values, and on the coefficient of determination (R2), which indicates how well a model fits the variability of the actual measures.
To answer our subsidiary research questions, we used a feature of Random Forests that computes each variable's importance for prediction accuracy.

KEY FINDINGS AND CONCLUSION
The sense of community model yielded an RMSE of 3.1 and a R2 of 76.5%. The three most predictive variables were: median age, percentage of people providing unpaid care and index of work accessibility. In the literature consulted, the most influential factors were percentage of married people, gender and age [5].

The participation model yielded an RMSE of 1.9 and an R2 of 62.5%. The three most predictive variables were: proportion of people in intermediate occupations, proportion of people with a level 4 or higher of education and proportion of small employers and own account workers. In the literature consulted, the most influential factors were age, ethnic fragmentation and level of education [4].

The lack of reliable national measures of sense of community and participation with a neighbourhood level breakdown prevented us from producing geographically detailed measures. However, the high accuracy of the models shows the feasibility of our approach and makes further research desirable. Dependent variables at LSOA level – according to the ONS geographies – would allow to train models applicable at neighbourhood level. The majority of the datasets used were from the 2011 Census. The Census is organised every ten years: to produce updated measures of CC dimensions, the most predictive variables should be collected more frequently.

The difference between the most predictive variables in our models and the indicators found in the literature may be an artefact of the different mathematical frameworks used by Random Forests and statistical methods. Further work from both data and social science perspectives is needed to better understand the causes of these differences.

REFERENCES

[1] BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16, 3 (2001), 199–231.

[2] FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S., and AMORIM, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The Journal of Machine Learning Research, 15(1), 3133-3181.

[3] GIBBON, M., LABONTE, R., and LAVERACK, G. (2002). Evaluating community capacity. Health & social care in the community, 10(6), 485-491.

[4] RUPASINGHA, A., GOETZ, S., and FRESHWATER, D. The production of social capital in US counties. The journal of socio-economics 35, 1 (2006), 83–101.

[5] SENGUPTA, N., LUYTEN, N., GREAVES, L., OSBORNE, D., ROBERTSON, A., ARMSTRONG, G., and SIBLEY, C. Sense of community in New Zealand neighbourhoods: A multi-level model predicting social capital. New Zealand Journal of Psychology 42, 1 (2013).

Dealing with Big Data in the Local Government

Vania **Sena**, University of Essex

In this paper, we explore the challenges of using Big data for local governments and present the activities of the ESRC Data Research Centre for Business and Local Government which has been recently established by the ESRC for the collection and curation of the data from both local governments and businesses.
To showcase the activities of the Centre, the paper will also focus on the work which has been carried out with the Colchester Borough Council to understand what drives the vitality of the high street.

Using ‘big data’ to inform local policy decisions

Hannah **Durrant** and Julie **Barnett**, University of Bath, UK

The way we understand the potential contribution that ‘big data’ can make to the policy making process is being shaped at the intersection of a number of interconnected agendas. Efforts by government at all levels to ensure policy is informed by rigorous and reliable evidence have been revitalised by the challenges of significant spending restriction, the impact of austerity and a changing demand for services. This is, in turn, reinvigorating debates about the nature of evidence and the role that different forms of knowledge can and should play in the policy process. The importance of robust information governance protocols and procedures has become ever more pertinent, as instances of highly visible data security breaches amplify public and professional perceptions of risk associated with sharing and linking data. At the same time open data is seen to offer a vital mechanism for enhancing the transparency and accountability of governments and of public policies. These agendas do not always sit in easy alignment with each other, and tension and disjunctions between them are realised acutely where the use of big data in the policy making process is practiced at the local level.

Since April 2014, Bath and North East Somerset (B&NES) Council, NHS B&NES Clinical Commissioning Group (CCG) and an interdisciplinary team at the University of Bath Institute for Policy Research (IPR) have been involved in a co-produced research project to explore the potential for connected data to inform citizen-focused local policy and practice. Funded by a Transformation Challenge Award, the principle aims of the project have been to create, pilot and evaluate a process to change the culture of information sharing across public services; develop mechanisms (technologies and processes) for safely linking data; and realise the benefits of big data by generating new insights into public needs to guide policy development. Starting from the premise that there remains a need for local policy making processes to be problem-led and theory informed, the approach has involved the co-definition of policy problems and the application of innovative techniques for analysing linked administrative data and aggregate area-level statistics to better understand local need and co-produce solutions. The partnership is founded on the principles of knowledge-exchange, to enable sharing of ideas and ensure that the learning from the project is genuinely co-owned.

This action-oriented approach has generated significant proof of concept. It has initiated local protocols and processes for safe sharing and linking of data, and produced high quality data analysis to inform the development of policy options and decision making. Specifically, linking administrative data with demographic, socio-economic and health data has produced a more comprehensive picture of the dynamics of financial hardship, and the use and effectiveness of public health services in the local area. It has also challenged established convictions about the extent to which current services meet public requirements and preferences. Alongside these benefits it has brought to the foreground two notable and linked issues associated with further advancing policy making in the big data era. A discussion of these issues forms the basis of this paper. They are:

1. significant and multi-dimensional professional concern regarding the practicalities and implication of sharing and linking data, and beyond this, of open data; and
2. the applicability of deploying data collected and retained for administering current services for other purposes.

Alongside a substantial appetite to exploit the potential of big data to improve the base of evidence on which policy draws, local policy making communities are expressing a number of reservations about the drive to such data-derived knowledge; both in practice and in principle. To some extent these reservations are associated with the challenge that such shifts present to professional expertise and the extent to which big data analytics are relegating professional knowledge to a lower order in the hierarchy of evidence. These reservations are not new or specific to the 'big data era'. However, they are resurfacing and coalescing with concerns about connecting multiple sources of data, and the implications for individual privacy. In a mixed economy of public service provision, the ethical consequences of drawing together data given to particular providers on explicit or implicit terms of use, and in relation to a perceived degree of independence from the state, requires particular consideration. These concerns are amplified by an abstract and decontextualized perception of public fear about data security that, in turn, constrains both the potential that new forms of data and data analytics might offer the policy making process, and the possibilities for public engagement with these approaches.

There are also significant and valid reasons why policy making communities at the local level are concerned about the use of data collected for one purpose to inform another (albeit related) purpose. There are considerable limitations to the use of existing – particularly administrative data – to inform new enquiries. Such data is primarily collected for the purposes of providing a current service and is thus understandably designed to be fit for this, and not broader objectives. The project, whilst recognising the proliferation of existing data, has identified a number of instances where the questions being asked require additional or different data to inform policy and practice that aims to meet changing needs. The challenge for policy-making in the big data era goes beyond technological and analytical developments, and is associated with how these new forms and modalities of data can inform process by bridging the gap between the questions for which we don't have data and the data for which we don't have questions.

Using data to protect the public

Andrew **Goodman**, The Home Office

Exploring government administrative data to hold governments accountable in the Big Data Era

Mihály **Fazekas**, University of Cambridge

The proposed workshop aims at informing conference attendees about the state-of-the art in using big data for measuring quality of government, corruption, and spending efficiency, with particular focus on data and indicators of government contracting, company ownership, political officeholders, and public sector budgets. It would allow attendees to have a unique glance into a recently starting cutting edge Horizon2020-funded research project led by the University of Cambridge (Department of Sociology: Dr Mihály Fazekas, and Computer Laboratory: Dr Eiko Yoneki) called DIGIWHIST, and to explore how the wealth of government administrative data can be used to hold governments accountable.

Public procurement, the focal point of DIGIWHIST, is the purchase by governments and state-owned enterprises of goods, services and works. It is one of the largest government spending activities in any country, representing on average up to 13% of GDP in OECD countries, and up to 29% of general government expenditure. At the same time, it is perceived to be the most corrupt government function across the globe ahead of justice or taxation. Even though governments throughout the globe, and across Europe in particular, are producing large amounts of administrative data describing public procurement contracts and tenders, this information is left largely unused for research and policy purposes up until now. Applying a Big Data approach strong in combining diverse data sources is expected to strengthen accountability and transparency of public administrations as well as unlock a whole new domain of research based on hard data rather than surveys of perceptions or self-reported crime.

Since public procurement is prone to corruption and budget deficit risks, high quality open data and innovative assessment tools are especially relevant for the efficient and transparent use of public resources. DIGIWHIST systematically collects, analyses, and broadly disseminates tender-level information on public procurement in 35 jurisdictions across Europe (EU28+). This data is linked to company and public organisation information on finances and ownership and to information on mechanisms that increase accountability of public officials such as asset declarations. The project uses innovative ICT-based measures and services which provide wide access to information about governments' spending and it involves private and public actors to actively collaborate in improving the quality and scope of the relevant data.

DIGIWHIST spans through the whole spectrum of activities from identifying datasets, collecting data, cleaning and combining diverse data sources, developing metrics which make sense of the immense richness of such a combined dataset, up until disseminating the results to policy makers, investigative journalists, and ordinary citizens in a tailor made form. Hence, the proposed workshop will be able to inform attendees of state-of-the art research methods and findings in the field.

The following activities are planned during the workshop, assuming sufficient time is available:

1. Introduction into data collection and cleaning methods
   Giving a unique insight into how data sources are identified, discussions with data holders are conducted, what the raw material for Big Data research is (i.e. semi-structured text files released by governments on a highly variable basis), and how data is extracted, cleaned, and linked. Diverse country examples will be given such as the UK, Hungary, Romania, and Italy.

2. Data scope and content
   Describing database content including variable list and variable content, and discussing in detail data scope (i.e. the universe of events covered by the database) as it differs per country. In addition, data scope is crucial as some of the missing and erroneous data points can be used as indicators on their own (e.g. failing to report criteria for assessing bids may indicate corruption risks in tendering).

3. Indicators
   A brief overview of the different indicator families developed will be provided such as indicators of transparency, corruption risks, and administrative quality. Then, a selected set of more innovative indicators will be discussed in

detail in order to demonstrate the strengths and weaknesses of the approach followed. For example, CRI (Corruption Risk Index), which is a composite indicator of corruption 'red flags' of the procurement tendering process, or PCI (Political Control Indicator), that assesses risks related directly to political officeholders owning or controlling government suppliers.

4. Joint exploration
   Preliminary versions of two crucial outputs of DIGIWHIST will be shared with participants and the opportunity for joint exploration provided. First, data and indicators on some countries will be released and a range of data analysis opportunities will be shown while attendees will also have the chance to explore them individually on their laptops (assuming technical background can be provided). Second, the Hungarian and Czech national web portals disseminating data and indicators will be presented live and functionality will be explored together with participants.

The workshop is designed to be interactive with a two-way dialogue at the centre stage. As the project will only have preliminary results to share, the participants will be invited to share their views and opinions about the results which can be taken into account when further developing the project.

The workshop will be conveyed by Dr Mihály Fazekas, but it will also make use of active contribution from the DIGIWHIST team, mainly from the University of Cambridge, but also involving researchers from our partner institutes such as the Open Knowledge Foundation.

Democracy and data: how data-driven policy making can avoid technocracy

Anthony **Zacharzewski**, Demsoc
Prabhat **Agarwal**, European Commission
Adam **Watson-Brown**, European Commission

Big data can visualise the complexity of the policy environment, but also challenges the policy making process. Data can give great insights, or lead to complacency and false certainty. Most of all, it is not enough just to have it you have to use it.
This session explores what a good policy process looks like in a big-data world.

Big data means that the policy process needs to be more agile. Agility in policy and implementation allows for corrective action and iterative interventions, informed by early monitoring of effects, better awareness of reality and continuous flow of evidence. Need to connect this into a process – show the discipline in the iterative process.

The potential prize is that it means that policy and the directives that flow from it could become more responsive, both to situational and geographical differences. Does that improve or hinder equality either of outcome or legal protection?

It also means that laws can be set with an intended effect in mind, and then repealed or amended once that goal has been met. Sunset triggers could be more effective than timed sunset clauses.

However, data-driven policy making must also be humble, democratic and ethically sound. Big data can lure us into a false sense that we know everything, and that we can solve anything using the right algorithm or the right user information. Open process needs open data alongside.

How can the political and participative elements of policy making sit alongside data?

How do we ensure that political values are exposed and transparent in a constructive way?

How do we ensure that agility is disciplined, transparent and constructive?

Online Social Network Privacy as a Collective Phenomenon

David **Garcia**, Emre **Sarigol** and Frank **Schweitzer**, ETH Zurich, Switzerland

Introduction

The problem of online privacy is often reduced to individual decisions to hide or reveal personal information in online social networks (OSNs). However, with the increasing use of OSNs, it becomes more important to understand the role of the social network in disclosing personal information that a user has not revealed voluntarily: How much of our private information do our friends disclose about us, and how much of our privacy is lost simply because of online social interaction? Without strong technical effort, an OSN may be able to exploit the assortativity of human private features, this way constructing shadow profiles with information that users chose not to share. Furthermore, because many users share their phone and email contact lists, this allows an OSN to create full shadow profiles for people who do not even have an account for that OSN.

Data

We use a publicly available dataset of Friendster, a former OSN with a very similar functionality as Facebook or MySpace. Before its social networking functionalities were discontinued, Friendster was crawled by the Internet Archive, leaving a snapshot of all the publicly available information at that moment. Our previous analysis of the connectivity patterns of the network [1] reveals that the first 20 Million users of Friendster were largely located in the US, before the OSN spread to other countries. This allows us to analyze these initial 20 Million users as a subset of users that live in the same country and create an account under similar conditions. Most users allowed their friendship lists to be publicly available, and some of them also let other users to see private features explicitly given by the user, such as age and gender. Within the subset we considered, 3,431,335 users had public profiles which were captured by the crawl, with personal information including birth date, gender, relationship status, and romantic interests. This subset contains a total of 11,074,009 undirected friendship links among these public profiles only. In addition, each user has an id number that indicates the order in which the user joined the social network, allowing us to analyze the growth of the network.

Methods

For each user in the dataset, we built a feature vector including their profile information, and different metrics of the distribution of features in their neighborhood at distances up to 3. We use this representation to make a data-driven simulation of the growth of the network, evaluating the capacity of the OSN to predict the sexual orientation of the users that did not join yet. The schema on the left of the Figure shows the full shadow profiles problem, which exploits the information of the users that joined the OSN up to time t and builds profiles for the non users based on the external links shared by a ratio ρ of the internal users. We base our prediction assessment on a random forest classifier, measuring its accuracy κ for users outside the network using only information and links shared by the users inside.

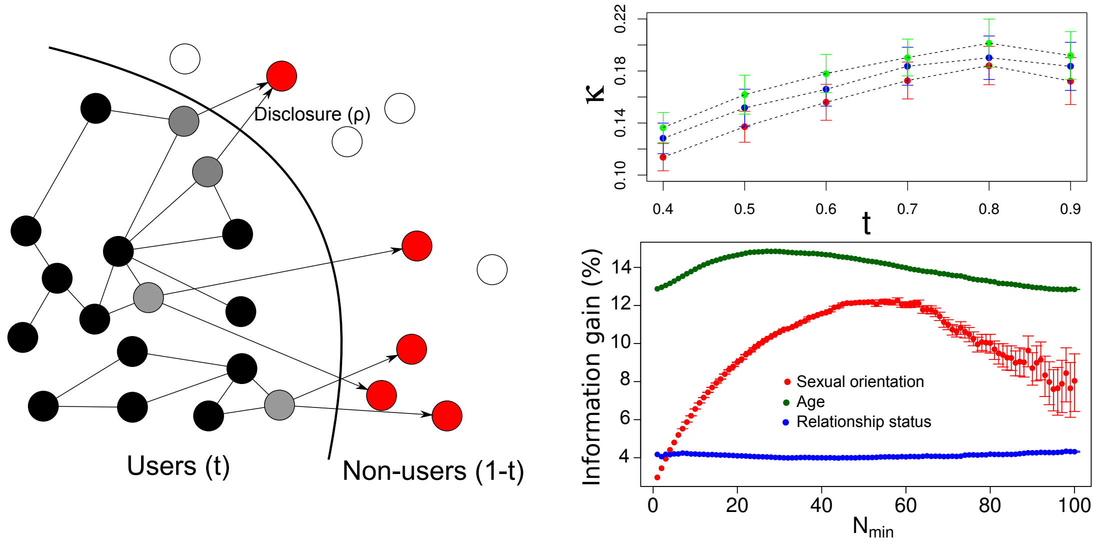


Figure 1: Left: Schema of the full shadow profiles problem. Right top: dependency of prediction quality over the growth of the network. Right bottom: Dependency of information content of private attributes with neighborhood size.

Results

Our first study [2] shows that prediction accuracy for sexual orientation increases with the size of the network, as shown if the right top of the figure. Furthermore, we computed privacy leak factors, as the regression results of the relation between private prediction accuracy and the tendency of others to share their information. We found positive and significant privacy leak factors, and a heterogeneity that makes certain sexual orientations more subject to this phenomenon. Our ongoing research extends this study to the features of age and marital status. Using entropy metrics, we quantify the amount of information about private attributes that is contained in the neighborhood of a user. The lower right panel of the figure shows the information gained by knowing the neighborhood of users with a minimum amount of friends in a social network. The high constant values of age show that this attribute is easy to predict for any neighborhood size, and the increasing trend of sexual orientation shows that the more friends a user has in a social network, the easier it is for that network to predict its sexual orientation.

Conclusion

While we do not provide evidence that shadow profiles exist at all, our results show that disclosing of private information is not restricted to an individual choice, but becomes a collective decision. In an interlinked community, an individual's privacy is a complex property, where it is in constant mutual relationship with the systemic properties and behavioral patterns of the community at large. We provided quantitative insights into the dependence of an individual's privacy to their respective community, and how far an OSN provider can utilize this dependency to create shadow profiles. Our work does not improve the methods to create shadow profiles; we limited ourselves to the application of existing methods to underline an already existing risk. We showed that, as the network grows and its members share their contact lists with the provider, the risk of privacy leakage increases. Given the fact that this dependency is present under generalized social interaction, we should consider privacy as a collective concept, where policies that rely purely on the control of individual users are not sufficient to preserve the right to privacy.

References

[1] David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. Social resilience in online communities: The autopsy of friendster. In Proceedings of the 1st Conference on Online Social Networks, 2013.

[2] Emre Sarigol, David Garcia, and Frank Schweitzer. Online privacy as a collective phenomenon. In Proceedings of the 2nd Conference on Online Social Networks, 2014.

Bodies of Data and the Person in Personalized Medicine

Jeffrey **Skopek**, Faculty of Law, University of Cambridge

It is often questioned whether there is anything that is truly exceptional, from an ethical or legal perspective, about big data. This paper will argue that there is—at least in the medical context. It will demonstrate that the use of big data in personalized medicine has the potential to transform this emerging field, giving rise to hard questions that cannot be adequately answered by reference to established principles of medical law and ethics.

The first transformation that I will identify and explore is epistemic: the use of big data in personalized medicine will bring about a fundamental shift in how patients and their medically relevant traits are known.

To understand nature and impact of this shift, however, one must first understand the current paradigm of personalized medicine. This paradigm is perhaps best exemplified by the field of pharmacogenomics, which seeks to understand the genomic differences within traditional disease categories—such as breast cancer—and use this knowledge to develop pharmaceuticals targeted to these differences. For example, after discovering that a protein known as HER-2 was over-expressed in 25-30% of women with breast cancer, researchers developed the drug Herceptin to target and prevent this over-expression.

Although this first generation of personalized medicine is still in its infancy, a second generation—fueled by developments in big data—can already be seen on the horizon. In the future, personalized medicine will not need to rely solely on a predictive model based on biomarkers and causal pathways that are well understood. Rather, doctors will be able to make even more powerful predictions on the basis of sophisticated algorithms crunching large-scale datasets that include every available type of data (genomic, environmental, historical, geographical, social, familial, etc.), the relevance of which is not yet known and may never be known. For example, an algorithm might predict that the drug Herceptin will be effective for a given breast cancer patient not because she has an over-expression of HER-2, but rather because she has 100 other traits that when combined are correlated with drug efficacy for reasons that we cannot explain. This possibility is, in my view, what makes big data transformative, and generative of hard and novel questions.

One of the core questions that I will explore in the paper is how this epistemic shift will impact the types of privacy concerns that are increasingly arising in the context of medical research. At first glance, it might seem that such concerns will only be aggravated by this shift. The idea here would be that research based on big data will involve amassing and sharing large-scale datasets, which would seem to increase the potential privacy harms caused by unauthorized access to the data. However, as I will argue, there is another way of thinking about the privacy implications of this form of personalized medicine.

To see the other side of the privacy story, one must recognize the ways in which the second generation of personalized medicine will (in its underlying epistemology) entail a shift from realism to operationalism—or in other words, a shift from a vision in which science seeks to provide us with knowledge of the world as it really is, to a vision in which it seeks to provide us with only operational knowledge. As applied in the medical context, this could entail a shift away from a world in which we try to obtain meaningful knowledge about the biological traits of the patient, to a world in which we merely try to make accurate predictions about them. In developing this account, I will demonstrate how the emergence of big data is thus not necessarily privacy-destroying, as many have suggested, and even might be privacy-protecting—at least according to traditional conceptions of privacy. I will also consider, however, whether this merely means that a new conception of privacy is needed.

The second transformative impact of big data that I will identify and explore concerns the types of data that will be included in medical research. In the era of big data, what constitutes medical information will no longer be confined to the information found in places such as clinical records, drug trials, and other medical contexts. Rather, big data analyses will reveal that other types of data—such consumer purchasing records, social media, etc.—can be just as important in predicting medical outcomes as the genetic or biological data gathered in the clinic. Thus, the use of big data will change how we understand the nature of personal medical data and where it can be found.

This shift will require that we rethink whether ethical standards that were developed to govern research using data gathered in the medical context should be applied with equal force to medical research using data gathered in other domains (e.g., Google's Flu Trends and Facebook's study of emotional contagion). The central question that interests me here is whether individuals should have the right to limit the use of their anonymized data in medical research. Against those who argue that such rights should be recognized on basis of autonomy, property, or other personal grounds, I will suggest that we should instead treat such data as part of a shared human biocommons. At the same time, however, I will identify and explore alternative, public-oriented rationales for granting such rights. Drawing on communitarian political theory, for example, I will suggest that we might grant limited rights of control in order to turn medical data into powerful tools of participatory democracy.

Challenges in Revealing the Bright Shadows of the Digitally Invisible

Justin **Longo**, Arizona State University; David M. **Hondula**, Arizona State University; Evan R. **Kuras**, Boston University; Erik W. **Johnston**, Arizona State University

The "big data" movement has quickly emerged in recent years and is poised to have a profound impact on policy making. As governance systems have greater access to more data, the ability to adapt policy interventions based on fine-grained evidence collected from ubiquitous sources will dramatically change the nature of policy formulation, implementation and evaluation. Big data accumulated through the online activity and social media engagement of individuals, electronic transactions, measurement by in situ and personal sensors, counters and smart meters, interactions with devices and control technology, and network-connected mobile technology can lead to important insights in policy areas relevant to human behavior. The accumulation of these data and associated metadata such as geolocation information and time and date stamps results in a previously unimaginable amount of data, measured with phenomenal precision, taken from multiple perspectives. Advances in data storage technologies now make it possible to preserve increasing amounts of data. And advances in data mining, analysis, and visualization technologies and techniques can yield valuable new insights.

But what happens in policy areas where the evidence collected fails to detect key targets that are invisible to the network of sensors, card readers, cell towers and servers? A new digital divide may be emerging in the presence of big data, with those on one side increasingly invisible when that data is used as the basis for policy analysis purposes (Haklay 2012). For those without a smartphone, without a bank account or credit card, living beneath and beyond the network of sensors, monitors and data capture points, their existence is being rendered increasingly invisible. As a result, policy making is blind to their existence and therefore does not reflect their reality.

One important population likely living outside the realm of big data is the homeless. While we have managed to make modern homelessness largely invisible through public policies and personal avoidance (Waldron 1991), big data is compounding those choices by rendering those living at the margins of our societies digitally invisible. Increasingly sensitive and precise policies are biased in favor of the digitally connected, though blind to the digitally invisible (boyd and Crawford 2011). We investigate how interventions can begin to lessen and bridge this big data digital divide, while protecting the privacy rights of the digitally invisible, addressing concerns about the surveillance state and engaging participants as partners in the research initiative.

Research Questions

- Is digitally invisible a real phenomenon, and if so what explains it?
- What challenges exist in engaging a homeless population in research supporting “big data”-informed policy making?
- How can the research participants collaborate in the co-production of the research such that they shape future initiatives and benefit from the findings?

Research Methods

This study is designed as a pilot project aimed at investigating procedural challenges in engaging as research partners a marginal population living in the shadows of the digital world. We focus on one measure of personal welfare amongst a specific population: individually experienced temperatures (IETs) amongst homeless individuals in Phoenix, Arizona (Harlan et al. 2008).

Extreme heat is among the leading weather-related health hazards across the globe. Strategies for minimizing heat-health risks are often based on fixed-point or place-based monitors, or population-level data. Our approach investigates the potential for revealing the challenges, preferences and behaviors of the digitally invisible through wearable temperature sensors carried by research participants. Technology-based initiatives to improve data reliability on homeless populations have centered on improving data captured by observers and interviewers (e.g., Morais 2014). Emerging technologies and data capture strategies make it possible to generate novel insights into urban environmental conditions and exposure at a very fine resolution - e.g., through crowdsourced temperature measurements from smartphones (Overeem et al. 2013). This trend suggests that heat-health response efforts and policies can become more targeted, tailored, and adaptive to match resources with the people, places, and situations where the need is greatest. To inform this movement, Kuras, Hondula and Brown-Saracino (2015) present an experimental approach in which research participants are equipped with a Thermochron iButton, a small and lightweight mobile sensor that measures and records instantaneous air temperature at 5-minute intervals. Participants clip their iButtons to a belt loop or bag such that the device is continuously exposed to the surrounding air as they go about their daily lives, and return the device to the study team after one week. This design has been implemented in two pilot studies to date enlisting over 100 individuals from diverse neighborhoods in Phoenix and Boston, Massachusetts.

This study will repeat this method with individuals living in the Phoenix area who self-identify as homeless, persons especially important to consider in the design of heat-health programs and policies as they disproportionately experience adverse health outcomes associated with exposure to hot weather (MCDPH 2014). We will ask ten participants to carry an iButton for two weeks in mid-spring 2015 (when temperatures in Phoenix are already approaching dangerous levels) and attempt to collect the device from them at the end of the study period. In addition, pre- and post-test interviews will be conducted with study participants to understand their perceptions and concerns with respect to privacy, obtrusiveness of the device and scope for more complex data capture protocols and greater researcher/participant collaboration. Past pilot studies with non-homeless individuals yielded low concerns about privacy and tracking using the iButton method, though we will explore this in greater depth in this study. The issue and measurement of IET is not the primary focus of this research - rather the procedure of distributing a data collection device to a homeless population, understanding the acceptability of and their experience in participating in the study, and collecting the device and learning more about the research participants' perspectives at the end of the study period are the main objectives.

Our paper will report on our findings with respect to the process of running this pilot experiment, speculate about opportunities to scale this experiment to other data collection devices and larger populations, and identify concerns about attempting to reveal the digitally invisible.

References

boyd, d. and Crawford, K. (2011). Six Provocations for Big Data. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Oxford Internet Institute, September 2011. Available at SSRN: http://ssrn.com/abstract=1926431

Haklay, M. (2012). Nobody wants to do council estates' – digital divide, spatial justice and outliers. Paper Session: 2121 Information Geographies: Online Power, Representation and Voice. New York: Association of American Geographers Annual Meeting. http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=44781

Harlan, S.L., A. Brazel, G.D., Jeanerette, N.S., Jones, L. Larsen, L. Prashad, W.L. Stefanov. (2008). In the Shade of Affluence: The Inequitable Distribution of the Urban Heat Island. Research in Social Problems and Public Policy. 15: 173-202

Kuras, E. R., Hondula, D. M., & Brown-Saracino, J. (2015). Heterogeneity in individually experienced temperatures (IETs) within an urban neighborhood: insights from a new approach to measuring heat exposure. International journal of biometeorology, 1-10.

Maricopa County Department of Public Health (MCDPH) (2014). Heat-Associated Deaths in Maricopa County, AZ. Final Report for 2013. May 7 2014. http://www.maricopa.gov/publichealth/Services/EPI/pdf/heat/2013annualreport.pdf

Morais, C. D. (2014). Using GIS to Identify Milwaukee's Homeless Population. GIS Lounge. December 11 2014. http://www.gislounge.com/using-gis-identify-milwaukees-homeless-population/

Overeem, A., R Robinson, J. C., Leijnse, H., Steeneveld, G. J., P Horn, B. K., & Uijlenhoet, R. (2013). Crowdsourcing urban air temperatures from smartphone battery temperatures. Geophysical Research Letters, 40(15), 4081-4085.

Waldron, J. (1991). Homelessness and the Issue of Freedom. UCLA Law Review, 39: 295-324

Digital Methods for Public Policy Research: Mapping Open Data as an Issue Online

Jonathan **Gray**, Richard **Rogers** and Liliana **Bounegru**, Digital Methods Initiative, University of Amsterdam, Netherlands

The digital age – and its abundance of new sources of data and computational methods – has significant implications for political and social research, as well as for how such research is being used in the service of public policy. Debates in this area revolve around methodological, theoretical, epistemological and ethical challenges that digital data and associated methods bring to research (Savage and Burrows, 2007; Burrows and Savage, 2014; boyd and Crawford, 2012).

In dialogue with these debates, several fields of research have been developed over the past years, from digital humanities (Berry, 2011) to computational social sciences (Lazer et al., 2009), as well as specific techniques or approaches within them, from cultural analytics (Manovich, 2011) to webometrics (Thelwall et al., 2005). A survey of these digital research practices can be found in Rogers (2014).

In this paper we introduce one such digital research approach developed at the crossroads between Sciences and Technology Studies (STS), Actor-Network Theory (ANT) and Internet Studies, namely digital methods (Marres & Rogers, 2005; Rogers, 2013; Venturini, 2010, 2012). In its essence, digital methods "refers to repurposing online devices and platforms (such as Google searches, Facebook and Wikipedia) for social and political research that would often have been otherwise improbable." (Rogers, 2014, p. 78).

To examine the potential applications of digital methods in the service of public policy research we take as a case study our recent work on mapping the concept of "open data" as a political issue online. Undertaken as a collaboration between the Digital Methods Initiative and the Institute for Information Law at the University of Amsterdam, the University of California, Berkeley and the global civil society organisation Open Knowledge, this research project asks: How can we trace the competing visions and values articulated around open data online? Who are the actors, what are the issues, and how are they related? And, more broadly, what can digital methods contribute to public policy research?

We take open data as a case study because within a remarkably short space of time the concept of “open data” has vaulted from being the rather rarefied preserve of a handful of information activists and technicians to possessing significant currency on the global political stage, featuring prominently in the speeches of Presidents, Prime Ministers, Mayors and Commissioners, as well as on the agendas of major international groups and organisations such as the G8, the G20, the OECD and the World Bank.

Advocates argue that the “open data revolution” will enable greater transparency, accountability and public participation; new civic applications and services; greater government efficiency; technological innovation and new businesses and startups (Gray, 2014; Kitchin, 2014). However, critics argue that open data initiatives may actually end up empowering the empowered (Gurstein, 2011) or acting as an instrument of a programme of austerity, neoliberalisation and marketisation of public services (Bates, 2012, 2013, 2014; Gray, 2014; Longo, 2011; Margetts, 2013).

In spite of its meteoric ascent and its extensive political implications and dimensions, open data remains largely under-studied as a political concept. We address this gap in research by using digital tools and methods to analyse and explore open data as a malleable and contested idea in several online spaces, including Wikipedia, the web and Twitter, and how these digital devices participate in the enactment of open data as an issue online. We find that the meaning of open data is continually reconfigured in response to shifting ideals, conceptions and practices of governance and democracy in different contexts and that different digital spaces enact different visions of this issue. Through hyperlink and resonance analysis, we find that open data is gaining significant traction amongst governments, major international institutions, large companies, media, and small group of “influencers” - but so far it seems to gain less traction amongst non-specialist civil society organisations and actors “outside the bubble”. Through the analysis of metadata on Wikipedia we find that open data is much more of a “digital commons” issue than an open government issue. Through network and co-hashtag analyses we find distinct groupings of actors promoting very different and competing visions of open data as a political idea – from governments promoting innovation and smart cities, to tech activists working on e-democracy or crisis mapping.

We conclude the paper with a number of recommendations for policy in the area of open data, as well as reflections on the theoretical and methodological opportunities and challenges for using digital “methods of the medium” in the service of public policy research, particularly for “born digital” or highly digitally mediated policy issues, including addressing questions of informational bias and representativeness online.

References

Bates, J. (2012). ‘This is what modern deregulation looks like’: Co-optation and contestation in the shaping of the UK's Open Government Data Initiative. The Journal of Community Informatics, Vol 8, No 2.

Bates, J. (2013) The Domestication of Open Government Data Advocacy in the United Kingdom: A Neo-Gramscian Analysis. Policy & Internet. Vol 5, Issue 1, 118-137. March 2013.

Bates, J. (2014) The Strategic Importance of Information Policy for the Contemporary Neoliberal State: The Case of Open Government Data in the United Kingdom. Government Information Quarterly. Vol. 31, Issue 3, 388-395.

Berry, D. (2011). The Computational Turn: Thinking About the Digital Humanities. Culture Machine 12: 1-22.

boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society, 1-18.

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. Big Data & Society, 1-6.

Gray, J. (2014) Towards a Genealogy of Open Data. Paper presented at the European Consortium for Political Research (ECPR) General Conference 2014, University of Glasgow.

Gurstein, M. B. (2011) Open data: Empowering the empowered or effective data use for everyone?. First Monday: 16:2-7.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., ... Van Alstyne, M. (2009). Computational Social Science. Science, 323, 721-723.

Longo, J. (2011). “#Opendata: Digital-Era Governance Thoroughbred or New Public Management Trojan Horse?” Public Policy & Governance Review. Vol. 2, No. 2, 38.

Kitchin, R. (2014) The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences. London: Sage.

Manovich, L. (2011). “Trending: The Promises and the Challenges of Big Social Data.” http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.

Margetts, H. (2013) “Data, Data Everywhere: Open Data versus Big Data in the Quest for Transparency”. In Bowles, N. Hamilton, J. T. & Levy, D. (eds), Transparency in Politics and the Media: Accountability and Open Government. London: I. B. Tauris & Co.

Marres, N. and R. Rogers (2005) “Recipe for tracing the fate of issues and their publics on the Web”. In B. Latour and P. Weibel (Eds.) Making Things Public: Atmospheres of Democracy. Cambridge, MA: MIT Press.

Rogers, R. (2013). Digital Methods. Cambridge, MA: MIT Press.

Rogers, R. (2014). Political Research in the Digital Age. International Public Policy Review, 73-87.

Savage, M., & Burrows, R. (2007). The Coming Crisis of Empirical Sociology. Sociology, 885-899.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. Annual Review of Information Science and Technology, 81-135.

Venturini, T. (2010). “Diving in magma: how to explore controversies with actor-network theory”. Public Understanding of Science, 19(3), 258–273.

Venturini, T (2012). “Building on faults: how to represent controversies with digital methods”. Public Understanding of Science, 21(7), 796-812.