Supplementary Appendix - Socialising Big Data: From concept to practice

Background

Socialising Big Data: Identifying the risks and vulnerabilities of data-objects was an Economic and Social Research Council (ESRC) funded project that took place from June 2013 to Sept 2014. It involved collaboration between social scientists from a range of backgrounds (sociology, anthropology, and science and technology studies), many of who were affiliated with the Centre for Research on Socio-Cultural Change (CRESC Manchester and The Open University) and the Centre for Economic and Social Aspects of Genomics (CESAGEN Lancaster), but also including other institutions.[1] The project aimed to advance the social scientific analysis of Big Data and digital practices to benefit academics, students, practitioners and policy makers. It did this by conducting three separate collaboratories with practitioners in turn from bioscience, national statistics and waste management. A final collaboratory brought together participants from all three. The project results were published in a CRESC working paper, *Socialising Big Data: From concept to practice* (available at: http://www.cresc.ac.uk/publications/working-papers/). This document is a Supplement to that working paper and consists of summaries of the presentations and discussions at each of the collaboratories, beginning with the final one.

[1] PI: Evelyn Ruppert, Goldsmiths, University of London. Co-Is: Penny Harvey, Manchester, CRESC; Celia Lury, Centre for Interdisciplinary Methodologies, Warwick; Adrian Mackenzie, Lancaster; Ruth McNally, Anglia Ruskin. Researchers: Stephanie Alice Baker, Goldsmiths, University of London; Yannis Kallianos and Camilla Lewis, University of Manchester, CRESC.

Appendix A: Summary of Final Collaboratory

The collaboratory begin with a provocative presentation on 'Big Data Claims' by one of the project team members, Adrian Mackenzie from the University of Lancaster. The rest of the collaboratory involved four sessions each focused on one of the four crosscutting themes – metrics, economies, ethics and collaboratories. Each started with a plenary presentation by a practitioner from one context followed by respondents from the other two. Participants were then divided into three breakout roundtables consisting of a mix of practitioners and project team members to discuss the theme and then report back to the group. The collaboratory ended with a plenary discussion of the collaboratory as a method and proposals for next steps.

Crosscutting Theme 1: Metrics

Plenary Presentation: Will Spooner, Eagle Genomics

Will Spooner discussed data metrics in genomic science. The field of genomics has been working with Big Data for about fifteen years and is therefore in a comparably advanced position compared to the other practical contexts. Genomic practitioners are interested in how they can *accurately* analyse data from integrated sources. Big Data is often defined in terms of the three Vs: volume, velocity and variety. Volume and velocity are becoming less of an issue and instead Will suggested that a fourth 'V' is becoming more important: veracity.

In the last decade, hybridisation technology was introduced which enabled analysis of specific parts – or the 'signature'– of the gene. This technology was highly successful but it was subject to instrument specific biases, which made comparing data from other studies increasingly difficult. In the last five years, new technologies have been introduced which are not subject to the same biases as the older instruments. Nevertheless, they introduce a new set of biases (more data does not mean no biases) when measuring the genome sequence. One of the problems of working with Big Data is that there is no 'truth set' and this givens rise to the issue of veracity. In light of this, genomic scientists are trying to discover new ways in which data can be integrated and remain accurate.

Big Data is systematic and holistic. The data at the scale that genomic scientists work with is systematic and covers an entire range (e.g. the whole genome). Big Data metrics reveal different facets or constitute a 'cut through' that range. Today, genomic scientists integrate the data so different stakeholders can make use of it (e.g. patients, clinical researchers, molecular biologists). In the field of genomics, Big Data can give stakeholders slices through these very large data sets. The notion of the boundary-object, as employed in the social sciences and outlined in the working paper, could provide a useful way to explore this further.

Respondent: Michail Skaliotis, Eurostat

- Big Data ecosystems require new ways of thinking about data and the roles of national statisticians. Within these ecosystems statisticians no longer design and produce data and need to think about and approach data in new ways. Practitioners in this field are still investigating what can be done with new forms of data beyond those produced by traditional methods.
- At present, national statisticians are facing enormous change. They must transform their existing infrastructures and practices at a time of budget restraints. One of the main challenges that lies ahead relates to data analysis – how can the vast quantities of data available be made into meaningful metrics and information? There are also many risks associated with predictive analytics. These issues require further debate among statisticians.
- What is required is clear communication of the risks and benefits of working with Big Data. The issues will differ depending on whether the audience is practitioners, health care operators or individuals. In the context of national statistics, for example, there is concern about how to communicate these risks and benefits to different stakeholder groups.
- Ethical codes are required for data analysis because of new security concerns. Responsibility is generally placed on users to secure their data. There needs to be more responsibility taken from IT specialists to strengthen security precautions and to limit the risks associated with data storage.
- It is generally assumed that genomic science is very precise, but this collaboratory has shown that genomic scientists still face many outstanding questions similar to those facing statisticians and waste practitioners.

Respondent: Joanna Hayduk, WRAP

- There is plenty of data in waste management but this data is not considered 'Big Data' per se. Practitioners are only beginning to think about how Big Data could be used in this domain. Given that local authorities are facing austerity cuts, one key question is whether Big Data could lead to efficiency gains by reducing costs.
- At present, tonnage is the most commonly accepted metric used by practitioners to measure waste. If the industry embraces new data sources, this will raise questions around alternative metrics (e.g. what units will they measure?).
- It is very challenging to create a consistent data set about waste in the UK. The 404 waste collection authorities across the UK all operate slightly differently (e.g. some collect waste in house while others contract out, different collection regimes etc.) and it can be difficult to develop consistent and comparable metrics.

Discussion

Group 1

- Group 1 discussed the quality and sustainability of new forms of data. For example, how could Big Data supplement existing data sources? Would Big Data involve the same metrics? Could it be used to save costs of data collection? Can it be used to increase recycling rates which in turn can improve scarce resource management?

- Big Data raises issues around certainty and trust. Existing approaches such as the analysis of time-series data are well established and trusted. Will new forms of data be trusted? How can we ensure that new data forms remain reliable? Would it be possible to make comparisons between Big Data sources and existing time-series data? Could they be integrated?
- Big Data doesn't always have a purpose. This makes it difficult to anticipate in advance what the missing metrics will be.
- This raises a fundamental question, what is a metric? A metric is generally understood as something standardised that counts units of some kind. Once a metric has been given a label how can it change? For example, there could be two metrics that measure exactly the same thing but are used for different purposes. Once you have a metric how difficult is it to change it? There could be opposition to change existing metrics that are well established (e.g. tons in waste management).

Group 2

- How do you determine what is relevant when deciding to join up different data sets? In practice, we do not have the luxury of endless analysis. Decisions have to be made at early stages of analysis such as deciding what data is 'important' and how it can be translated into meaningful information. More data is not necessarily the solution. The right data is key and the challenge is identifying this upfront and then making decisions about where to invest resources.
- There has been a move from prescriptive to predictive analytics. Practitioners in all three fields (genomics, national statistics, waste management) must thus think ahead and be prepared to make adjustments as one goes along and things change. Incremental and iterative experiments provide one way of managing this relationship.
- Currently, there are no mechanisms for pooling examples of experiments and analysing them collectively. They tend to happen on a smaller and large-scale sharing is not happening.
- We are moving potentially, in the next five to ten years, to an internet of things beyond our imagination in light of rapid technological change. As a consequence, redundancy needs to be built into systems. New sensors, online devices, wearable devices will be generating data in ways we cannot imagine. The generation of data in everyday lives is pervasive; many aspects of lives are now made into data. What will that mean in the future of Big Data?
- Veracity – the reliability and verifiability of data - will likely become more important and those sources that cannot demonstrate this will likely not be taken up.

Group 3

- New metrics are emerging. For example, national statisticians are trialling new (real-time) methods such as using social media platforms (e.g. Twitter and Facebook) to collect data. One of the advantages of these sources is that they provide good data about young people. Smart metres in homes also provide data about things such as occupancy rates.
- An example of how Big Data is currently being analysed involves the scraping of data from websites in order to calculate the rate of inflation by comparing a particular number of

products weekly. To improve accuracy, measurements can be taken at different intervals and compared.

- One of the key characteristics of Big Data is that it can be re-purposed and integrated with other data sets. However, working with Big Data in this way raises issues around data ownership and consent. When is/should consent be given? Is it possible for individuals to negotiate how and where their data travels?

Crosscutting Theme 2: Economies

Plenary Presentation: Barteld Braaksma, Statistics Netherlands

Barteld Braaksma discussed the potential of scarcity to drive economies of Big Data in an environment that is facing severe budget cuts. However, careful decisions must be made about where to target resources especially in the face a lot of 'hype' around Big Data. In spite of its potential, there is considerable uncertainty about the results of analyses and this makes securing future investment difficult. Processing and analysing Big Data also requires new technology and new technical capacities. Within the climate of uncertainty the challenge is thus how to prioritise resources to devote to Big Data. This is a common situation in the commercial world but the public sector is generally more risk adverse.

Statisticians were traditionally the only practitioners who could conduct large-scale surveys of populations, but now the situation has been reversed. Statisticians do not own Big Data sources. In the past they were the 'haves' and now they have become the 'have nots'. At the same time, collecting data in new ways (rather than asking people through surveys) could be more time efficient and cheaper and thereby address the challenge of resource scarcity. There is a strong feeling that these changes cannot be done by individual statistical offices alone but require collaboration as in the field of genomics and waste management. This includes public-private partnerships, which involve sensitive negotiations. But one difficulty in building such partnerships is that the private sector has different business models and funding streams. Finally there are also 'image risks' and issues of perception involved in using some Big Data sources generated by private technology industries.

Respondent: Gurdeep Sagoo, PHG Foundation

- Genomic research includes a number of different sectors and stakeholder groups that have different goals and interests. For example, in the private sector discussions tend to revolve around profit whereas, in academia, the onus is on knowledge creation with efficiency gains seen as a by-product. But the mixture generally does tend to work well in that the goals are often complementary. In genomics, the goal of using Big Data should be to improve the health of the population. The biggest challenge centres on funding and resources, which are relatively scarce.
- Over the past 5 years there has been a lot of discussion around drug development and 'personalised medicine'. The belief within medicine is that the doctor/patient relationship is about providing a personalised service rather than focusing on budget constraints. You treat

the patient in front of you rather than worrying about having the money to treat the patient coming to see you next week. There has been much work done recently on drug development, which can be more effective and minimise harmful reactions. Drug companies are investing a lot of resources in these areas of 'personalised' medicine.

- With respect to economies, the emphasis is on balancing the health needs of all of the population within a relatively fixed and limited budget. There is scope within the government system to pay for new drugs if there was a pandemic (e.g. the recent Ebola outbreak), but action requires political support. The 100,000 Genomes Project, which has been backed by British Prime Minister David Cameron, is a good example of a public-private partnership. The emphasis is on both improving health and economic benefits. It will be interesting to see how these health and economic benefits will be realised in the future.

Respondent: Steven Rose, Strategic Research, Birmingham City Council

- In the future, waste is likely to be viewed as a resource and based on an economic model where 'raw materials' must be 'harvested' from people's bins. This shift in thinking will have a significant impact on how data is collected and analysed.
- There is a lot of data available to local authorities. This could be used by the private sector to design technologies such as the GPS systems for fleet and waste vehicles.
- In the field of waste management, discussions focussing on recycling and waste reduction tend to be evangelical. The emphasis is on narratives about the common good and the public 'pulling together'. But individuals all have different needs and behaviours. They ask 'what's in it for me'? One way would be to make data available to the public so that individuals could make better decisions about reducing waste. Alternatively, consumers could use waste data to push companies to change their packaging materials. This kind of data could be used for lobbying and moving sustainability agendas forward.

Discussion

Group 1

- There is growing consensus that Big Data will benefit the country, but what will the actual benefits be? There needs to be a return on investments because training people in these new areas is expensive.
- 'Data science' as a discipline involves the skills and knowledge of a number of different practitioners, including statisticians, data analysts, and information technologists.
- New outputs and insights into Big Data must be flexible and reactive. If the government wants to know something new (e.g. the number of people on zero-hour contracts), Big Data could discover this information more quickly than standard practices.
- This begs the question, 'If you had an infinite amount of money to spend on Big Data, what would you do with it?'

Group 2

- What is the Big Data business model – what is the case to be made and its justification – for profit, for benefit, for public service? It is not merely about money, but other types of value (e.g. environmental, public goods and health).
- Much of the rhetoric about Big Data is predicated on its promise, it is being sold for its potential, that upfront investment in prevention could lead to long term savings and that in the future benefits will be derived from working with this data. For example, in genomics there is a promise of what it might deliver but the pay-off keeps getting postponed. In official statistics the momentum is based on the promise that society will benefit in the long term. There is also the assumption that there is a risk of missing something important if Big Data is not used.
- You are not generating data organically if you don't have a platform. The platform is an important part of the ecology of working with Big Data.
- Consider the economies of social media platforms – they depend on advertising, which is a business model that is being standardised.
- Different relations exist between different actors. The public sector has an interest in commercial data and wants to access it so they can change their working practices. It is not often acknowledged that businesses also have a strong interest in public data (e.g. census data). The relationships and dependencies go in many directions.
- There are regional differences in how the public sector operates. Some countries have detailed information about the commercial sector and others do not. The question of economies is complicated by these different circumstances.

Group 3

- Working with Big Data can create new markets. For example, the pharmaceutical industry's use of genomics data.
- One motivation behind using Big Data could be to acquire a competitive advantage (e.g. a new contract in waste management). The length of these contracts (often 25 years or more) makes it difficult to adopt new data practices once a public-private partnership has been established. Many questions emerge around data ownership especially since different partners have different requirements.
- Data is a commodity. Why are we paying for people (in the private sector) to get rid of our waste when they are profiting from these materials?
- Questions were raised around intellectual property. National statistics data are more open and transparent while commercial partners have an interest in patenting their data, which is likely to prevent others from using their data.
- The re-use of data is part of the economies of Big Data but some re-uses are less desirable such as in decisions on health and car insurance that penalise particular groups and can lead to inequalities. Compare this to collectivised uses targeted at better services and efficiency gains.

- It is unclear what personal data is being used for and the rules about data protection are unclear. For example, data may be shared on crowdsourcing or patient sites. Questions emerge around how this data is used and by whom.

Crosscutting Theme 3: Ethics

Plenary Presentation: John Bland, Greater Manchester Waste Disposal Authority (GMWDA)

John Bland spoke about the ethics of working with Big Data in relation to waste management. Big Data could potentially revolutionise waste management practices through improved services and efficiency gains. Whereas standard practice has traditionally been to collect information around tonnages, improving waste recycling requires more personalised services and individual data (i.e. so that barriers - whether physical such as the supply of the right "bins", or behavioural such as people who can't be bothered - which stop residents participating are able to be more effectively addressed). This move towards integrating Big Data into waste management could facilitate behavioural change through peer pressure, penalties, credits or rewards. At the same time, working with Big Data raises ethical issues around privacy and data protection (e.g., in relation to the Data Protection Act).

The determination of what constitutes ethical action is not always clear, particularly when it applies to pressing social issues, such as, climate change. Moreover, what are the consequences of using this information if it implicates people from certain socioeconomic or ethnic groups? For example, if a particular ethnic community produces more waste is it ethical to disclose this information to the public? Is it ethical to assimilate data to make generalisations about certain demographic groups? Is it ethical to charge people for residual waste or to reward them for recycling? These questions must be framed not only in relation to improved services and efficiency gains but assume particular significance when applied to the public good.

Respondent: Pete Brodie, Office for National Statistics (ONS)

- Pete Brodie responded by discussing some of the ethical issues around using Big Data. These concerns were canvassed in relation to current research on Big Data at the Office for National Statistics (ONS).
- Pete noted two types of uses of data: crowd (general) and individual (targeted). Tesco's loyalty card scheme is a case in point. This data can be used in several ways, to target sales to individuals via special offers or to generally improve services by configuring stores to appeal to their local demographic. The former generally raise more concerns about ethics though the latter is still profit oriented. People's attitude towards Big Data will thus depend largely upon how it is used.
- The same applies to the context of waste management. Ethical issues emerge around data usage. Is Big Data being used to penalise people or to improve their lives via better services? The public is likely to react negatively if Big Data is used to uncover the contents of their bins, whereas crowd information could be used to encourage people to recycle at the

local level via rewards, peer pressure and community engagement. These examples demonstrate the importance of considering how Big Data will be used as this is likely to shape people's attitudes and behaviour.

- The ONS currently run four projects on Big Data involving mobile phone data, Twitter data, Smart Meter data and Price Information data collected online. These different modes of data collection each raise particular ethical challenges. When accessing mobile phone or Twitter data, for example, the ONS has to consider ethical issues regarding surveillance, privacy and informed consent. While data of this kind can be readily accessed, people have not necessarily been informed that it will be used for commercial or research purposes and might not be willing to have organisations use this information. The issue of informed consent is particularly important to the ONS because it is a trusted organisation.

Respondent: Gurdeep Sagoo, PHG Foundation

- Gurdeep Sagoo responded to John's presentation by discussing the ethics of working with Big Data as a genomic practitioner. Within the field of genomics, there is a long history of thinking about the ethical, legal and social issues around data. The Ethical, Legal and Social Implications (ELSI) Research Program was established in 1990 as an integral part of the Human Genome Project (HGP) to foster basic and applied research on the ethical, legal and social implications of genetic and genomic research for individuals, families and communities.
- In 2014, the Realising Genomics (RG) project focused more specifically on the ethical, legal and social issues (ELSI) raised by the implementation of WGS/WES in clinical practice. The issues associated with the clinical implementation of next generation sequencing (including WGS and whole exome sequencing (WES)) are highly relevant to the UK government's 100,000 Genomes Project. The report recommended using targeted testing, when possible. Coincidental or incidental findings raise ethical questions. In undertaking this project, the most contentious ethical issues that arose were when individual rights come into conflict with greater public and societal interests. Questions arose, for example, over whether patients should be able to opt out of receiving incidental findings when they are clinically relevant. For example, should patients be able to limit the extent to which their sequence data is shared with other researchers, commercial providers and the NHS? If so, to what extent should this happen? These questions led to recommendations around informed consent and increased transparency. Gurdeep then asked whether informed consent is ever really possible? There are significant barriers to acquiring informed consent, particularly as it applies to logistics around time and cost.
- In genomics there is also a blurring between research and clinical practice. In research, anonymising and aggregating data can help to overcome some of these ethical challenges. There is, however, a concern that anonymity can be lost through data linkage. If genomic data is linked to other NHS databases, as the 100,000 Genomes Project seeks to achieve, then it could become relatively easy to de-anonymise genomic sequence data. This could lead to individuals being discriminated against on the basis of their genomic sequence.
- These ethical issues encourage us to think seriously about whether health is an individual or societal issue, and who should take responsibility for our wellbeing.

Discussion

Group 1

- Group 1 commenced by reflecting on the ethical issues around public services, in particular the tension between individual interests and collective benefits.
- This led to discussion about the distinction between the public and the private. The distinction between the citizen as a participant in the public sphere as opposed to a consumer interacting with private corporations is becoming blurred at a time when citizens are being asked to pay for public services.
- Many ethical issues are derived from the uncertainty around how individual data will be used.
- Working with Big Data raises issues about data ownership (e.g. who owns the data, the individual or the data generators?) and rights. Is there such a thing as digital human rights? Is data an extension of the person and, if so, is the use of it in certain contexts a violation of the person?
- There is a tension between the ethical practices of companies and academics as exemplified by the controversy over the ethics of Facebook's recent study with academic researchers that involved manipulation of people's emotions without their knowledge or consent.

Group 2

- There was discussion around the ethical implications of data linkage and consent, as well as the practicality of offering anonymity or the right to withdraw data. From a technical standpoint, this might be impractical. Instead, what is required is political discussion about the risks and vulnerabilities around data integration, particularly as this applies to health.
- The preoccupation with individual privacy risks overshadowing other important ethical issues. For example, searching for correlations in Big Data sets can lead to unexpected results that have the potential to stigmatise certain ethnic groups.
- From an ethical standpoint, there is a growing need to consider the costs and benefits of working with Big Data. If these ethical issues are not communicated clearly to the public, the potential benefits of Big Data can be lost.

Group 3

- Working with Big Data raises ethical issues around trust. There is a need for people to know how their data is going to be used in order to mitigate potential ethical dilemmas. Perceptions and fears around Big Data need to be addressed and distinguished from myths about how data is used, stored and collected.
- There was debate over whether data ownership is an ethical or legal issue. At present, the law is not equipped to deal with the ethical concerns associated with Big Data.
- In order to assess adequately the costs and benefits of working with Big Data, there needs to be an emphasis on assessing the potential risks involved.
- There is a growing need to identify and communicate the overall value of Big Data rather than merely focusing on the negative risks of sharing data.

Crosscutting Theme 4: Collaboratory

Plenary Presentation: Steven Vale, United Nations Economic Commission for Europe (UNECE)

Steven Vale spoke about collaboratories as a method to engage with different communities of practice, providing reflections on the collaborative approach and recommendations on how the method could be developed in the future. Over the past two years Big Data has become an emerging topic within the field of national statistics. Yet, despite growing interest in Big Data, at present there is not much understanding. Working for UNECE, Steven agreed to attend because of his interest in Big Data and because his team is trying to introduce new ways of working with Big Data in official statistics. In this regard, there are similarities between collaboratories as a mode of engagement and the working practices of the UNECE, which brings together volunteers with shared interests to a common forum to discuss, share, explore and arrive at common solutions. In both cases, practitioners commence with a blank canvass from which they brainstorm and address key questions they need to solve in order to develop a common vocabulary – a particularly important task when assembling people from diverse contexts and regions. A key part of the UNECE's work revolves around how to proceed and innovate. This begs the question, what are the next steps after the collaboratory? Steve emphasised the value of considering these questions to continue the momentum developed over the past year.

Despite these similarities, collaboratories also differ from the working practices of the UNECE in significant ways. The UNECE's activities are situated within a relatively specialised community. Steve found it interesting to be exposed to novel perspectives, methodologies and metrics from different practitioner groups. He suggested that collaborations of this kind create the opportunity for interdisciplinary engagement. For example, the practices of genomic scientists or waste management practitioners could inform the production of national statistics in new and exciting ways. Moreover, much of the UNECE's work has a global dimension. Whereas the collaboratories were mainly based in the UK, it could be interesting to adopt a more international dimension through virtual communication or other means. The key issue is to be flexible and willing to learn, realising that findings may differ from what was initially expected. An example of such a mode of collaborative exchange is the Big Data Sandbox, which provides a technical platform and a collaborative environment where statistical organisations can play with Big Data, and experiment with new tools and methods. This type of experimentation could be trialled in future collaboratories.

Overall, Steve found the collaboratories a useful approach. Facilitation and structure is key. Getting the structure right is a challenge. There needs to be enough structure to stimulate meaningful exchange while leaving room for creativity and debate. Collaboratories are most effective when practitioners are at similar stage of development as is typically the case with Big

Data, which raises common issues around access, reliability and ethics across the different practical contexts: genomics, national statistics and waste management. In sum, collaboratories provide a very useful approach of cross-cultural engagement. It would be beneficial to find a mechanism to continue these discussions (via an electronic forum, for example) so as not to lose the momentum established.

Respondent: Will Spooner, Eagle Genomics

Following from Steven's evaluation of the collaborative approach, Will Spooner reflected on how successful the collaboratories have been. He discussed the issues that arose and how the team might have done things differently. Will remained unclear about what a collaboratory is and how the approach differs from other forms of engagement, such as, focus groups, networking events and workshops where people assemble to address specific topics and problems. He proposed that collaboratories are distinguished from other kinds of collaborative research because they examine how things in the world are constituted as objects (e.g. Big Data as a Digital Data Object (DDO)). In this regard, collaboratories could be said to disrupt existing hierarchies and to challenge the status quo.

The central aim of the Final Collaboratory was to advance the analysis of Big Data practices. For Will, this appeared to represent a social science point of view. Within the field of genomics, the focus is on establishing new definitions about Big Data and novel ways to express and classify Big Data that recognises its social context rather than merely focusing on its technical qualities. The social context of Big Data is vital. What is it about the social and cultural context of Big Data that makes it "BIG"? Will suggested that one way to measure the success of a collaboratory is in terms of its *usefulness* to practitioners and the community it seeks to inform – the findings must be useful to the community and practitioner groups by providing insight into a specific subject area and the ability to influence practice and policy. In this regard, Will asked how we intend to influence policy?

Will attended the collaboratories in order to acquire a social scientific understanding about Big Data. Working for Eagle Genomics, a company that uses open source software and data that is emerging from academic communities, Will needs to understand the real value of open data in sociological terms. This enables formulating the value proposition of their services to sponsors and customers through a common vocabulary and to innovate new services that can increase their value. Big Data has to be understood in the context of disruptive innovation – it provides a new way of working that opens up new markets that companies such as Eagle Genomics aim to target. Will found the Final Collaboratory more useful than the first because it exposed him to different perspectives and practitioner groups, which led to new insights and ways of looking at Big Data. He found the working paper particularly useful because it synthesised the findings and provided the sociological terminology and vocabulary that he originally sought. For example,

the notion of Big Data as a "boundary object" demonstrated how the value of Big Data might be increased. For Will, this Final Collaboratory led to the growing realisation that Big Data is inherently social compared to other sorts of data. The value of Big Data is maximised by integrating multiple data sources to arrive at a multidimensional viewpoint of the problem at hand. Big Data is characterised as a data object that has relevance to multiple stakeholders beyond the use for which it was originally collected (e.g. Facebook or Twitter data).

Respondent: Celia Lury, Centre for Interdisciplinary Methods, University of Warwick

In evaluating the collaborative approach, Celia Lury commenced by reflecting on the different pronunciations of the term, "collaboratory", which she suggested reflects something of the history of the term. Whereas the term "co-laboratory" emerged in a scientific context, the "collaboratory" as a collaborative method is more widely used in the humanities and social sciences. Celia contended that both inflections are useful. Employing the collaborative approach in relation to Big Data has shaped how the team organised the collaboratories in terms of the kinds of questions asked and the practitioners with whom we collaborated. Interdisciplinarity is often precipitated by a notion of crisis, the idea that there are pressing problems that require disciplines to come together. Big Data is an emerging field that disrupts and challenges standard working practices and lends itself to interdisciplinarity and asking questions such as: What is Big Data as a problem space and how can we this space through different modes of collaboration? Big Data involves a redistribution of data collection and research methods expertise and the restructuring of infrastructures, which necessitate engagements with a wider range of collaborators. In order to address questions around the social life of Big Data then requires engagement with practitioners from both the public and private sector.

From a social science perspective, collaboratories provide a testing ground for concept development. It is important to consider whether we have learnt anything about the kind of "socialising" involved. For example, what are the frameworks for thinking about Big Data? In terms of policy, we have legal, economic and political frameworks for thinking about Big Data. Should we add a social framework for thinking about Big Data and, if so, how would a social framing be different from these existing modes of analysis? From this perspective, collaboration may be thought of as an iterative process distributed not only in terms of space, but time. In terms of knowledge production, collaboratories bring social scientists into the collaborative process from the outset rather than merely being there to challenge, critique and problematise the findings of social scientific research. What is exciting about collaboratories is that they help us to move beyond individualised disciplines and projects by providing a method to develop and tests concepts.

Respondent: Hannah Knox, Dept. of Anthropology, University College London

Hannah Knox responded to these discussions by reflecting on the genesis of the project. She noted that the collaboratories were conceived at CRESC as a way to make academic research more useful and to have a greater impact. The method was designed as an experiment to trial the impact of opening up communication by assembling people (researchers, stakeholders and practitioners) at the initial stage of questioning and agenda setting rather than merely documenting findings towards the end of a project, as is typically the case in academic research.

In light of these objectives, Hannah emphasised the interrelationship between collaboratory as method and the topic of Big Data. When problematised, Big Data requires particular forms of collaboration between different stakeholders and practitioners. Despite the will and ambition for collaboration, commercial and political interests can act as powerful boundaries to collaboration on the topic of Big Data. This Final Collaboratory provided a neutral space in which to discuss some of these challenges, such as attempts to integrate Nectar card data from Sainsbury's loyalty card schemes with that of other organisations, which was blocked due to Sainsbury's existing relationships with other commercial enterprises. In this regard, collaboration provides a useful way to understand the problems of working with Big Data. Through this approach, for example, we can identify who the important players are and ask questions about this burgeoning topic. The Final Collaboratory has revealed some of the key players in the field, but certain stakeholders were absent, such as, the users and producers of Big Data.

Hannah ended her presentation by thinking about how to proceed with the collaborative approach. She emphasised the value of developing a shared vocabulary, but was curious about whether this would take an oral or written form (via publications or an extension of the working paper, for example). She then asked whether collaboratories would lead to new modes of experimentation or novel research projects, concluding by highlighting the importance of talking collectively about the benefits of collaboration as a method.

General closing discussion

The group engaged in a general discussion and raised the following points about the collaboratories and what was accomplished.

- What has been started here should not sit on a shelf; this was just a beginning.
- One of the outcomes has been the establishment of a diverse network of people engaged in questions of Big Data. Out of this we could consider possibilities such as a project involving waste management authorities, ONS and social scientists.
- The project has widened horizons and enabled connections that might not otherwise have happened. The diverse and conversational approach of the final collaboratory was

appreciated; it enabled people to speak without the fetters of 'credentials' and provided a safe environment to think out loud. That said, more provocation and controversy could have been introduced.

- The working paper was especially helpful. But an alternative approach to the structure of the collaboratories would be good to consider. The position of the social scientists seemed to be more as observers rather than active participants. It would be good to consider a model that is more of a mix.
- The insights of the project need to come forward especially in the face of documents such as the EC data driven economy – why not think about a data driven society?
- More private sector involvement would be a good next step as well as from data scientists, privacy groups, data journalists and so on. Additional follow-up actions would be good to identify.
- It is good to talk about Big Data but what is also needed is a space for not just flying ideas but doing Big Data that could support the move to policy development.
- The concept of socialising is useful for understanding the different norms of different disciplines and interests and the benefits of mixing or 'socialising' them.
- How might the international aspects of Big Data be better leveraged? Recognizing that Big Data generated by online platforms cuts across national borders it would be useful to have forums that address this.

Appendix B: Summaries of Context-specific Collaboratories

Collaboratory 1: Metrics for DNA

How to make sense of what is going on in contemporary genomics?

Overview

Genomics is possibly one of the most sustained attempts at counting things ever undertaken in the modern life sciences. Genomics and various related 'omics' fields have developed an intricate infrastructure for generating, storing, sorting out and counting data on living things.

This workshop brought together genomic researchers, social scientists, various experts and stakeholders to discuss the metrics for genomic data in various contexts ranging across its making, its circulation (through databases and other infrastructures), and its analysis and use in various settings (academic, scientific, clinical, government and commercial).

A core question was: What kinds of metrics best serve and account for genomics as a counting project'?

There is a veritable deluge of data metrics associated with genomic data: base pairs/genome, cost/base pair, runs/day, Tb/day, Gb/genome, $000s/genome, 15 months 'doubling time', as well as all the numbers and counts used in accounts of genomics in practice (18,000 genes; 3 billion base pairs; 2% of the genome; 35 population groups; 23% missing heritability, 100,000 genomes).

These metrics often drive and motivate developments in genomics. Yet at the same time, many discussions, interventions and presentations on genomics suggest an acute awareness of missing metrics; of the need to develop metrics that are more refined, accurate or relevant.

When and for whom and in what contexts are basic metrics useful (e.g. like cost/base pair)? We are interested in finding out about what really counts in genomics today and why.

This collaboratory was organised by Professor Adrian Mackenzie and Dr Ruth McNally and held on 2 December 2013 at the Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK.

Speakers included:

1. **Neil Hall**, Advanced Genomics, University of Liverpool
2. **Sarah Ayling**, BBSRC Genome Analysis Centre (TGAC)
3. **Lucy Raymond**, Cambridge University
4. **Gurdeep Sagoo**, Public Health Genomics Foundation
5. **Laura Clarke**, Re-sequencing Informatics, European Bioinformatics Institute (EBI)
6. **Chris Hayman**, Amazon Web Services
7. **Will Spooner**, Eagle Genomics

8. **Rasko Leinonen**, European Nucleotide Archive (ENA), European Bioinformatics Institute (EBI)

Some background readings on Big Data and Genomics were circulated in advance:

- Boyle, J. (2013). Biology must develop its own big-data systems. *Nature*. 499.
- Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeño-Tárraga, A., Cleland, I.,...& Zalunin, V. (2013). Facing growth in the European nucleotide archive. *Nucleic Acids Research*, *41*(D1), D30-D35.
- Hall, N. (2013). After the gold rush. *Genome Biology*, *14*(5), 115.
- Kodama, Y., Shumway, M., & Leinonen, R. (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Research*, *40*(D1), D54-D56.
- Piton, A., Redin, C., & Mandel, J. L. (2013). XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *The American Journal of Human Genetics*, *93*(2), 368-383.
- Wetterstrand, K. A. (2013). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *URL Available at: www. genome. gov/sequencingcosts*.
- Wright, D. C. (2011). Next steps in the sequence: the implications of whole genome sequencing for health in the UK. PHG Foundation.

Summary of Presentations

Neil Hall: Genomics and economics in a research setting

Neil Hall discussed the relationship between genomics and economics in a research setting. DNA sequencing is used to measure a range of factors from sequencing genomes, assaying differences between genomes, mutation screening, population genetics and environmental sequencing. The metrics that scientists are most interested in are the length and quality of the sequence. Sequencing costs are crucial for science. Rapid developments in DNA sequencing technologies have revolutionised contemporary genomics. In just 10 years, the cost to sequence a genome has reduced from $100m to $5k. Improvements in the time and cost it takes to sequence a human genome have led to new opportunities and efficiency gains (e.g. single molecule sequencing provides new opportunities for real-time research).

Genomic sequencing has become a commodity with a predictable market value. But although the cost to sequence a genome has become relatively predictable, project complexity remains difficult to quantify and measure. Sequence reads and quality scores require more tertiary analysis to be made meaningful (e.g. via multi-sample processing, data aggregation and population structure analysis). Applications and platforms, moreover, can produce substantially different results. Part of the problem is that the analytical tools and gene sequencing devices designed to read and process genomic information have different properties, errors, costs and

read lengths. These result in methodological differences in terms of content, quality and scale. Another metric that cannot be measured is the time is will take to analyse data. The complexity of a given project will largely depend on what data is produced and where the data leads, which cannot be known beforehand. Different analyses require different data properties. We are not simply measuring genomes, but making them meaningful. While increased throughput and decreasing costs place pressure on analyses, for some projects long read technologies are prioritised over cost. In short, the current status of biotechnology is a compromise between cost, quality and length.

Sarah Ayling: Wheat Genome Assembly Strategies

Sarah Ayling discussed the different approaches taken to assemble the bread wheat genome: a large, highly repetitive and polyploid genome. The most common metric used to describe assembly quality is N50, which captures data about the fragmentation of the assembly. However, this metric neglects data related to the accuracy or completeness of what has been assembled. The quality assessment tool (KAT) developed at BBSRC Genome Analysis Centre (TGAC) is based on kmer-spectra, which can be used to reveal properties of the genome sample, in addition to properties of the resulting assemblies and their completeness. Other metrics regularly used include genome/assembly size, ploidy, sequence read coverage, number of contigs, number of mapped transcripts and their sequence similarity (percentage identity), and cost.

Working with complex plant genomes and large genomes can raise a series of challenges. When working with genomic data, the problem is that the metrics you want to know before conducting your sequencing experiment are often missing. Researchers do not always know the ploidy, heterozygosity levels or genome size, for example, before they start sequencing. In effect, this means that a degree of pilot sequencing is always required to get an idea of the scope and complexity required for the project design.

Lucy Raymond: How to make sense of DNA variants as a cause of disease

Lucy Raymond discussed how to make sense of DNA variants as a cause of disease. Big data technologies process data into small, specific and true variants for the individual so that they can understand their disease and its implications to their wider family, improve their treatment and access to appropriate health care. Big data has the potential to help researchers to find the single variant that is the cause of disease. But the greater the number of variants in the data itself, the more difficult data is to analyse. Issues arise, for example, over which method(s) should be used to identify the true sequence of the DNA. Complexity poses a potential problem for analysis because there are many variables to measure. Paradoxically, the more knowledge that emerges about any one rare variant, the less certain we seem to be that this information is sufficient to reveal or predict the disease severity.

At present, we cannot measure the cause of the disease. Just because a variant is detected in an individual with a disease, this correlation does not imply causation. When looking for the important variants associated with intellectual disability, for example, many factors can be at play (e.g. genetically highly heterogeneous, multiple possible causes). Processing this data into meaningful information consequently introduces a series of risks, which include balancing judgement of variant causation with the risk of harm if the data is incorrect or misinterpreted.

At present, the key metrics that clinicians rely on daily vary from the quality assurance of gene analysis, gene variation, quantity, the cost of testing, and whether the variant has been reported before. In the long term, key metrics will include cost reduction, improved clinical utility, and the potential for cost effective individualised health care. To improve the quality of analysis we need a greater volume of genomic data pertaining to patients with rare disease.

Gurdeep Sagoo, Public Health Genomics Foundation: Evaluating the (costs and) benefits of NGS

Gurdeep Sagoo evaluated the "costs" and benefits of next generation sequencing (NGS). In the context of health care, economic evaluation is used to identify, measure, value and then compare both the costs and the consequences of two or more alternative programmes or interventions in order to make a recommendation on which alternative should be used. Initiatives of this kind aim to accelerate the application of advances in biomedicine and genomics within health services for patient benefit in ways that are fast, effective, fair and responsible.

The application of economic evaluation in the field of the genetic testing using next generation sequencing technology has raised a series of challenges. This is because many different metrics are needed for such an evaluation. These can be epidemiological in nature (prevalence) or analytical in nature (measures of sensitivity and specificity related to both assay and tests). While economic evaluation can introduce additional metrics such as cost-effectiveness, the current challenges remain mostly methodological. Health economists must develop new approaches or reach a consensus on the appropriate existing methodology to use in order to overcome challenges related to which analytical approach to use, which costs and outcomes should be measured, and how these should be valued. At present, next generation sequencing technology raises more questions than answers.

Laura Clarke: Metrics for high throughput genomics projects

Laura Clarke presented on metrics for high throughput genomics projects. Large scale data collection projects, such as the 1000 Genomes Project, Blueprint and Hipsi, collect metrics about their data which fall into 4 main categories: quantity, quality, identity and consistency.

A primary challenge for these projects is that they meet their targets in terms of quantity. This includes whether the appropriate number of experiments on the correct number of samples are conducted. The 1000 Genomes Project overcame this issue by tracking the data that the different sequencing centers produced and declaring data freezes to track how much data was available on each sample and to determine whether they met the specified targets.

In addition to quantity, the quality of data plays an important role in deciding whether the sufficient number of experiments have been completed. In the Blueprint Epigenome project, for example, expression as measured by RNA-Seq experiments undergo quality checks by comparing the alignment of the sequence to gene models and ensuring that a sufficient proportion of the sequence falls in gene bodies. Identity and Consistency are also checked to ensure that samples and data are internally consistent. These require more specialised methods that are created for, and then applied to, the project in question.

Chris Hayman: Research Computing on Amazon Web Services

Chris Hayman, from Amazon Web Services (AWS) provided a commercial perspective on genomics and cloud computing. Big data technologies offer new possibilities to the field of genomics. New DNA sequencing devices have the potential to introduce efficiency gains by reducing the time and cost of recording the series of bases that encode genetic information.

In spite of the opportunities that Big Data introduce, these technologies raise a series of problems regarding the diversity, complexity and volume of data produced by gene sequencers. The challenge that research labs experience is largely concerned with volume and variety, namely how to manage, analyse and disseminate the increasingly large data sets generated by genome sequencing platforms.

The complexity of genomic research has introduced further challenges, such as data management and access to valuable data sets. Previously, large data sets (e.g. The Human Genome Project (HGP)) proved difficult to locate, download, customise and analyse. Amazon Web Services has devised a solution to this problem, introducing new technologies (e.g. AWS Cloud) that provide scalable, cost-effective, flexible and secure storage services (e.g. Amazon S3) to manage and analyse genomic data (e.g. Amazon Kinesis & EMR). In addition to hosting data, this service also provides access to a variety of public data sets.

Will Spooner: Measuring Commercial Bioinformatics

Will Spooner, the founder of Eagle Genomics (2008), put forward a commercial perspective on genomics R&D. Big data has conventionally been defined in terms of the "3Vs": Volume, Velocity and Variety. Spooner considered this definition inadequate. He proposed an emerging fourth variable metric: Veracity. This move towards veracity as an integral (missing) metric for genomic data was traced through a history of the field of genomics. From a commercial

standpoint, customers were initially focused on volume, namely how to consume, store and manage the large data sets and resources available. Customers then became interested in their own genomic data, and the increasing rate at which it was being generated. This occasioned a shift towards velocity, from processing their data on workstations to high performance compute clusters or cloud computing technologies. Once volume and velocity were measured, users began to extract value from the data, which required a variety of data sources (e.g. integration both with the consensus reference, and also with data assets collected by the enterprise in previous R&D efforts). More recently, the emphasis has been on a new metric: veracity, as users seek to assess the quality of "raw" data.

Assessing the veracity of genomic data raises methodological issues concerned with processing bioinformatics (i.e. how to store, organise, retrieve, and analyse biological data), and ethical concerns regarding how diagnostics developed from "omics" technologies should be regulated in the clinic? At present, there is much debate in this area, which represents the current frontier of modern translational genomics.

Rasko Leinonen: Organisation and growth of NGS data in the European Nucleotide Archive

Rasko Leinonen discussed the organisation and growth of next generation sequencing (NGS) data in the European Nucleotide Archive. NGS platforms are producing data with significantly higher throughput and lower cost. A portion of this capacity is devoted to individual and community scientific projects. As these projects reach publication, "raw" sequencing data sets are submitted into next generation sequence data archives, such as the European Nucleotide Archive (ENA).

The ENA is a public data archive for nucleotide sequences operated as part of the International Nucleotide Sequence Database Collaboration (INSDC) with NCBI and DDBJ. The ENA has grown out of the EMBL Data Library, which was first released in 1982. ENA measures the growth of the next-generation sequence data using multiple metrics. The metrics most important to the ENA concern the number of sequenced samples and sequencing studies (e.g. the cumulative count and doubling time of studies), the number of sequenced reads and bases, and the number of sequenced genomes and transcriptomes. These metrics are then grouped by the INSDC Archive (ENA, NCBI, DDBJ), submitters and users (geographic region). The data is used by both internal and external user groups.

The proliferation of data produced (e.g. the volume of data and rate of increase) poses a potential problem for the ENA in terms data management and storage. A key challenge is how to choose the appropriate technology for data persistence based on number of identifiable objects, type and volume of data. The rate of growth must be monitored to assess future sustainability of current technologies, to assist in choice and adoption of new technologies, and

to factor in the upper limit to identify and store data (e.g. POSIX file systems are not convenient for storing large number of files). Large volumes of NGS data are best compressed and stored as files outside any database.

Collaboratory 2: What Counts? Big Data and Official Statistics

Official statistics is one of the most comprehensive attempts to provide meaningful information to governments and decision makers. National Statistical Institutes (NSIs) have developed rigorous methods for collecting and analysing data to produce and publish statistics related to the economy, population and society.

The collaboratory brought together national statisticians and social scientists to explore the range of meanings and implications of Big Data for official statistics by attending to the question: *What counts when using Big Data sources to produce official statistics?*

Big Data sources represent both an opportunity and challenge to official statistics. While Big Data have the potential to introduce efficiency gains (e.g. improved timeliness, cost savings), they raise significant methodological and ethical considerations regarding data quality, protection and management. From a methodological standpoint, working with Big Data sources requires new technical skills, infrastructures and capacities. In this collaboratory, we focussed on how these issues require rethinking about what constitutes an official statistic. How does Big Data remake the substance and meaning of what is measured and captured, and how does this come to inform what counts? What counts also raises the question of who does the counting: what organisation or authority organises and mediates the making of 'official' statistics, with what tools, methods and consequences?

These questions were addressed with regard to Big Data-related projects and initiatives currently taking place within NSIs (e.g., mobile phone data, social media sentiment analysis, Google trends, etc.). Participants included the Socialising Big Data project team and statisticians from England, Estonia, Ireland, Netherlands, Eurostat and the UNECE. The collaboratory was organised by Dr Evelyn Ruppert and held 3 - 4 February 2014 at Centre for Creative Collaboration, London, UK.

Speakers included:

- **Barteld Braaksma**, Statistics Netherlands (CBS)
- **John Dunne**, Central Statistical Office (CSO), Ireland
- **Pete Brodie**, Office for National Statistics (ONS)
- **Jorrit Zwijnenburg**, Statistics Netherlands (CBS)
- Margus Tiru, Positium
- **Kaja Sõstra**, Statistics Estonia (SE)
- **Susan Williams**, Office for National Statistics (ONS)
- Michail Skaliotis, Eurostat
- **Steven Vale**, UN Economic Commission for Europe (UNECE)

Some background readings on Big Data and Official Statistics were circulated in advance:

- Karlberg, M. and M. Skaliotis (2013). Big Data for Official Statistics – Strategies and Some Initial European Applications. *Conference of European Statisticians*. Geneva, Switzerland: 1-10.
- United Nations Economic Commission for Europe (2013). Big Data, Big Impact? *Conference of European Statisticians*. Geneva, Switzerland: 1-9.
- United Nations Economic and Social Council (2013). Big Data and Modernisation of Statistical Systems. *Report of the Secretary-General*, United Nations: 1-29.
- United Nations Economic Commission (2013). What does "Big Data" mean for Official Statistics? *Conference of European Statisticians*. St Petersburg, Russia: 1-9.

Summary of Presentations

Barteld Braaksma: From research to preparation for implementation

Barteld Braaksma discussed several Big Data initiatives taking place at Statistics Netherlands (CBS). Big Data sources (e.g. transaction data, scanner data, traffic sensors, satellite photos, mobile phone data) provide new opportunities to measure mobility, economic and social activity. These methods tend to be timelier and more frequent than previous approaches, but have been applied with various rates of success. One of the major Big Data initiatives at CBS is using social media to monitor public sentiment. Digital methods (e.g. Sentiment Analysis) can be used to extract sentiment from social media messages and analysed according to statistical themes (e.g. economy, education, media). These results can then be used as sentiment indicators to reproduce and/or compare to survey measures of consumer confidence.

In order to produce high-quality information, statistical processes must be governed by sound methodologies. Statistical outputs must be relevant, accurate, reliable, timely, accessible, coherent, comparable and consistent. In this regard, working with Big Data raises a series of familiar and non-familiar methodological challenges from access to privacy, data analysis and management. From a methodological standpoint, Big Data poses a threat to traditional survey models of classification and implementation. It is difficult, for example, to compare Big Data to survey data because these approaches measure different things (behavior vs. responding, often dealing with devices not people). Big Data sources and analysis tools reconfigure what we measure and value. What counts, and how this should be classified, is increasingly a product of the data available for analysis. Classification systems must continue to be relevant to be implemented in the future or else they will lose much of their value. So a key methodological priority is how to reconcile this new approach with statisticians' established desire for long time series and stable, coherent classifications systems.

Big Data means that statisticians no longer have a monopoly over data collection and management. They must work with data suppliers, which raises confidentiality and privacy issues, and acquire new technical skills, which is challenging in an environment undergoing severe budget cuts. Big Data is relatively quick to process and accessible at an unprecedented

scale. The challenge is how to 'cut through the noise' and overcome continuity issues by producing reliable and relevant information for future statistical analysis. In sum, despite the enthusiasm for Big Data research, there is reluctance among statisticians to implement Big Data into official statistics due to concerns about traditional quality measures. To solve these issues, statisticians will need to establish new partnerships outside of statistics and work with data owners, researchers and different user groups.

John Dunne: Big Data coming soon…to an NSI near you

John Dunne, from the Central Statistics Office (CSO), discussed some of the opportunities and challenges of working with Big Data sources. NSIs have a privileged place in legislation and are perfectly positioned to access and combine data sources for statistical purposes. Despite this power, working with Big Data raises a series of challenges that correspond to those associated with administrative data sources (e.g. data harvesting, linkage, classification, representation, privacy). Connecting data sources has proved particularly problematic. The CSO have responded to this issue by establishing a National Data Infrastructure (NDI). The primary differences between Big Data and administrative data relates to scale and ownership (owners typically require a licence to operate, legal constraints on data linkage). Large data sets are not necessarily more reliable.

Traditionally, statistics have been defined by the questions statisticians ask (e.g. survey model). Big Data, conversely, is configured more by the methods available than preconceived questions or systems of classification. Increasingly, it is the data that drives the questions rather than the other way around. This represents a significant change from existing statistical practices. Instead of trying to impose classification models onto Big Data sources, statisticians should focus on clustering and distinguishing patterns from these new data sources in meaningful ways. Big Data sources (e.g. call detail records, mobile positioning data, transaction data, electricity consumption data) mean that statisticians are able to ask new questions of data that were previously considered impossible. Electricity Smart Meter Data, for example, could be used to estimate household composition based on electricity consumption patterns. One of the benefits of this new approach is that Big Data tends to be livelier and cost efficient than traditional statistical methods. The fact that Big Data sources are often transnational in scope also presents new opportunities for collaboration.

Given the sensitive nature of Big Data, and the (passive) means by which much data is collected, NSIs must demonstrate responsible statistical leadership when working with these data sources by establishing common data protection principles and practices – 'just because you can, doesn't mean you should' use data sources. Processing Big Data also requires new statistical and technical capabilities, as opposed to management skills. It is difficult to justify building new technical infrastructures when the value of these data sources is yet to be specified.

Outsourcing and downsampling may assist in overcoming these technical challenges, but working with third parties raises significant privacy and confidentiality concerns, both perceived and actual (addressing these concerns represents a further challenge).

Big Data is changing the role of national statistics as information is being collected and processed at an unprecedented speed and scale. The future of NSIs is not threatened, but statistical organisations will need to evolve and adapt by developing the necessary capabilities to work with Big Data sources. Each data source (e.g. Big Data, administrative data and open data) has its own characteristics and inherent value. Rather than focusing on definitions, statisticians should seek to combine this data to enhance their value and eliminate unnecessary burden on respondents. One way to achieve this is by making use of existing data flows. Statistical agencies must continue to align their practices with the UN Fundamental Principles of Statistics (particularly, Principle 5: choose data sources with regard to quality, timeliness, costs and the burden on respondents) and demonstrate responsible statistical leadership.

Pete Brodie: ONS data strategy – How administrative data and Big Data fit together

Pete Brodie, from the Office for National Statistics (ONS), discussed the parallels between working with Big Data and administrative data, both of which form a crucial part of the ONS's data strategy (e.g. Vision of 2020). There is much uncertainty around what constitutes Big Data. The difference between Big Data and administrative data is arguably a matter of volume. The practices of working with large data sets involve issues of data linkage, data mining, repurposing, and warehousing. The kinds of issues that arise when linking administrative data for statistical purposes correspond to those associated with linking Big Data sources (e.g. reuse, repurposing, representativeness, cleaning, classification, comparability, continuity, confidentiality and privacy).

The increasing use of administrative data sources for statistical purposes introduces challenges around quality measures. Assessing the quality of data sources is crucial to statistical analysis given that the objective of NSIs is to produce statistics to inform good government and evidence based decisions. In contrast to survey collected data, administrative data tends to provide better coverage (i.e. it is more like a census rather than a survey), but is generally less timely (e.g. small businesses may only provide VAT records annually) and produces poorer metadata. Big Data should be more timely, accessible, and cost effective than traditional methods. At the same time, Big Data is more prone to change than administrative data (e.g. will Twitter still be popular in 10 years?), and its methods and metadata are less transparent. Census data, by contrast, is relatively small in volume. The volume of Big Data sets could potentially pose major storage issues (data warehousing could possibly solve this problem). Big Data sources are not governed by the same quality measures as surveys. Quality needs to be measured in a different way. Those working with Big Data need to focus on modelling errors

rather than standard errors, as is typically the case with surveys. The ONS has addressed these issues by producing 23 basic quality indicators for statistics involving administrative data. These indicators can also be applied to Big Data sources.

The key challenges for Big Data centre around issues of data quality, comparability and storage. Using social media will produce statistics based on social behaviour rather than economic data. Rather than seeking to reproduce survey quality measures or outputs, quality should be assessed according to whether Big Data is useful for decision-making purposes and policy evaluation given that this is the purpose of official statistics. Big Data could be used for auxiliary purposes to provide quality checks on survey data. Methods could be used to improve and complement rather than replace each other (e.g. a Bayesian dual frame approach). This is particularly important when working with Big Data sources given that users' habits can change (e.g. location information on mobile phones, preventing access etc.). Big Data also raises questions around privacy. Among NSIs legislative differences exist. In contrast to Ireland, legal access is currently a major issue in the UK. The ONS have established an Innovation Lab to test new methods on Big Data in secure and original ways. Data warehousing will be a key priority in the future given that much internal data at the ONS is 'siloed' (processed and stored in different systems). In light of these challenges, the ONS is focusing on three areas of development: data integration and aggregation, data collection (i.e. seek to reduce burden on respondents, administrative data to replace survey data where possible, moves toward electronic data collection as the norm), and data storage.

Jorrit Zwijnenburg: The use of Big Data in Statistics on Employment and Wages

Jorrit Zwijnenburg, from Statistics Netherlands (CBS), discussed the use of administrative data statistics on employment and wages (in this presentation, the terms administrative data and Big Data were used interchangeably). CBS compile statistical outputs on employment, wages and hours. These are published on a quarterly and annual basis. These outputs have different quality measures. Quarterly outputs measure employment and wages (e.g. used to calculate GDP). The emphasis is on timeliness, as compared to annual outputs, which provide much more detailed levels and structures of employment, wages and hours and measure more variables (e.g. regions, gender etc.). Annual outputs include huge mirco datasets at job levels. This information is used by researchers to link employment figures to information on education and health. Big Data provides the opportunity to measure monthly outputs. There are multiple uses of this data from tax authorities to insurance agencies, pension funds, and security schemes.

Administrative data is collected from several different sources (e.g. Tax Authorities and the Employee Insurance Agency). Processing this data can be challenging because it involves data linkage (e.g. to the Business Register), imputation, editing and analysis before the results can be disseminated. Special attention is needed for completeness, relevance, imputation methods for

missing data, seasonal patterns, and distinguishing real from artificial changes. Despite changing the input source from survey to administrative data, the processing tools that CBS use for analysis have remained the same. More efficient tools are currently being developed to deal with Big Data sources.

When compared to previous methods, Big Data introduces a series of benefits from better coverage, higher frequency (data received weekly), improved timeliness and efficiency gains. At the same time, working with Big Data generates practical and methodological challenges from missing variables to comparability issues. Statisticians must also develop new technical infrastructures and capabilities to process this data. In light of these challenges, there is hesitation among statisticians to shift to Big Data.

While Big Data may assist statistical processes and improve timeliness, its quality and use differs from traditional approaches and further experimentation is required before being adopted by statisticians. Incomplete data sets and associated issues of quality, comparability and availability (e.g. open data) raise questions about what should count as an official statistic (e.g. “raw” data?) and highlights the need for NSIs to reassess their role as statistical agencies. How NSIs use data sources must be considered in relation to user needs (i.e. different users will have different priorities and quality measures), given that the role of these organisations is to provide statistics for society and decision-making purposes.

Margus Tiru: Feasibility study on the Mobile Positioning Data for Tourism Statistics

Margus Tiru explored the possibilities of using mobile positioning data (MPD) – information about the positioning of mobile phones stored by mobile network providers (i.e. mostly call data records, data downloads and location updates) – to measure tourism flows and movement patterns. MPD is different from existing data sources in that it requires new technologies and methods. Despite raising privacy concerns, this data is highly valuable to statisticians, researchers, state and commercial organisations, such as the tourism industry, because in tracking their digital geographical footprints it reveals the behaviour of individuals (e.g. what destination someone has visited and for what duration) and can therefore be used for risk assessment, investment and marketing purposes. In terms of volume, mobile positioning data is relatively small but it requires significant analysis and computational power. It is processing, rather than storing this data, that makes it challenging to work with.

While MPD covers tourism data well (e.g. inbound/outbound trips, duration, destination), important indicators are missing (e.g. accommodation, mode of transportation, purpose of the trip, expenditure, socio-demographic breakdown). Assumptions can be made, but there are problems with deducing correlations (e.g. commuters working weekends, children using parents’ phone). It would be beneficial to measure transaction data, but data of this kind is highly sensitive.

MPD provides many benefits for NSIs from new indicators to improved timeliness and better coverage, but large quantitative samples also introduce methodological challenges from data processing to data linkage/harmonising, coverage issues, comparability, data ownership and management. These methods, moreover, are not always reliable or cost efficient. Working with MPD also raises ethical issues associated with legislation and privacy, even if the data is used with 'good' intentions. These issues highlight the need to establish a legal framework for official statistics to obtain data in legal, cost-efficient ways.

Susan Williams: Internet Search queries within migration statistics

Susan Williams, from the Office for National Statistics (ONS), discussed the potential of Big Data to inform statistical processes in general and to measure population statistics in particular. In one ONS Big Data project, Internet search queries were used to measure migration to the UK using Google Trends: a public web tool that gives a search volume index (SVI) showing how often a particular search-term is entered on Google search relative to the total Google search volume in various time periods and regions. The results were promising with Google search queries containing the term 'polski', for example, corresponding to official statistics from Labour Force (LF) survey estimates of Polish nationality.

While preliminary studies suggest that these data could provide insight into emerging events (e.g. population densities and flows, flu outbreaks), in contrast to survey models, Internet search queries provide very little information about their users (e.g. age, sex, nationality), which limits statistical analysis. Instead, the platform configures what type of data can be measured and is made available for analysis. Although the device provided relatively accurate results regarding larger EU8 populations new to UK (i.e. predominately young adults, who speak foreign languages and use the Internet), it was not as effective for measuring other populations (e.g. some time series and populations were too small to track).

Moreover, despite being timely, accessible and free, working with the device raises a series of methodological challenges from understanding how data is produced to discontinuity issues (e.g. changing APIs), which raise significant questions around reliability, consistency and representation. Despite obtaining good results when matched against LF statistics, within the ONS there is only the appetite to use the device for quality assuring official statistics.

Working with commercial providers to access sensitive information raises further concerns over privacy and legislation, confidentiality, monetisation and access. Those working with these data sources must also be aware that users could modify their behaviour to avoid detection (e.g. opt out, rejecting cookies, usage of Google search), which would impact the future success of these methods.

Internet search tools may be able to improve and inform statistical processes by providing quality assurance around existing statistics and indicators about population statistics (e.g. identifying tourists and migrant populations before they enter the UK). While the potential impact of Big Data for national statistics is promising, at present it is hindered by the lack of available methodology, quality measures and access to underlying data. To move forward, statisticians need to collaborate with commercial providers to achieve better access and learn more about their methods (e.g. Google).

Kaja Sõstra: Big Data in Statistics Estonia

Kaja Sõstra, from Statistics Estonia (SE), discussed some of the issues that arise when working with Big Data in relation to initiatives currently taking place within SE. Big Data forms an important part of SE's data strategy (2013-7). Existing data sources are mostly based on surveys. They tend to be extremely burdensome, costly and have a low response rate. Data collection in contemporary statistics, by contrast, is mainly electronic. One of the benefits of working with new data sources is that they can be used to produce relevant and timely results that accord with SE's objectives. By using resources effectively, Big Data technologies can also result in efficiency gains (e.g. data linkage, rapid collection, avoiding supplementary input work). In this regard, Big Data has the potential to solve some of the problems associated with producing official statistics. However, in order to process Big Data, statisticians require new technical skills and resources which is challenging in a context undergoing severe budget cuts and governed by compulsory EU regulations around output (e.g. regulations around microdata can cause problems when working with administrative data) and quality measures. SE have responded to these issues by developing and implementing innovative methods to process this data.

Big Data could be used to solve some of the problems associated with producing official statistics. These issues were canvassed in relation to several case studies.

- In 2013, SE completed the first phase of developing a methodology for a register-based Population and Housing Census. Data was collected mostly from administrative registers. There were several problems with this approach, a major issue being that respondents' registered place of residence was inconsistent with their real place of residence for about twenty per cent of the Estonian population. Big Data sources, such as mobile positioning data, could potentially be used to complement these methods by estimating the Estonian population during the census.
- Working with secondary data sources also raises methodological issues around classification. When developing methods to measure Energy Statistics in the household sector, for example, statistics need to conform to certain categories (e.g. cooking, lighting, space and water heating). Smart meter data could possibly solve this issue by modelling electricity end use.

- The surveys that produce statistics on wages and salaries raise similar practical issues. Data collection tends to be burdensome and survey methods can produce inconsistent results (i.e. employers report similar data to different authorities, missing variables, different reporting times). One way to resolve these issues could be to combine tax and statistical data used for producing labour cost index statistics.
- In a similar way, transaction data could be used to reduce burden on respondents, improve response rates and quality measures, potentially replacing the Household Budget Survey (HBS), in part or altogether.

In conclusion, Big Data sources could be used to complement existing statistics by providing more cost efficient, detailed statistics and better timeliness. In order to implement these changes, however, extra-resources are required to develop robust methodologies (practices and software). The key challenge for NSIs is how to work with Big Data sources in a way that is cost-effective, consistent with statistical principles (impartiality, reliability, relevancy, profitability, confidentiality, transparency), and complies with intl. methods and regulations.

Michail Skaliotis: Closing Observations

Michail Skaliotis, from Eurostat, provided some closing observations on the collaboratory. The shift toward Big Data is still in the early stages of development and must be seen as an evolution in order to assess the current position of NSIs. Implementing Big Data raises significant methodological and economic challenges (budget and resource constraints), especially in a conservative environment where there is ambivalence about its value and outcomes. How best to train data scientists is an issue that needs to be considered. Despite these concerns, there is some momentum. Constraints can be overcome through building new partnerships and international collaborations with various experts and stakeholders (e.g. discussions about creating a European Big Data Analytic Service). Collaborations of this kind will enable statisticians to learn from the experience of others and reduce investment costs.

Small statistical offices cannot compete with large third party data sources. Internet companies produce different daily and hourly estimates addressing similar socio-economic phenomena as statistical agencies. Perhaps the future of official statistics is to play another role: providing accreditation or certification as a statistical authority on what measures to use (e.g. what is the best inflation estimate). In sum, reflecting on the tremendous changes that have occurred in the last four years (2010-2014), there is reason to be very optimistic about the future of national statistics.

Steven Vale: Closing Observations

Steven Vale, from the UN Economic Commission for Europe (UNECE), provided some closing observations on the collaboratory.

- He began by questioning the value of the term 'Big Data' suggesting that 'big' was not the key characteristic of the data sources they seek to describe. He proposed using the term 'new data' instead, but conceded that it raised similar problems of relativity (i.e. what's big/new now will appear small/old in the future).
- The term 'Secondary Data Sources' (data not collected first hand by statistical organisations) was proposed to describe these new data sources given that almost all the strategic and technical issues associated with Big Data apply to administrative data. Thus, the utility of making distinctions between Big Data and administrative data sources was questioned.
- Big Data raises significant privacy concerns. Much research is focused on how to manage these issues, both perceived (e.g. exacerbated by the NSA scandal/Snowden revelations) and actual. In reality most statistical offices neither have the resources, nor the interest to trace specific people. Instead, they are interested in discovering trends and aggregates. This needs to be communicated to the public.
- Big Data raises methodological challenges regarding classification, definitions and comparability. What is being measured is not commensurate with existing methods; data is increasingly driven by the technologies available. Perhaps it is a mistake to impose preconceived measurements and modes of classification onto Big Data sources. The alternative is to identify what's new about this data and to discover what they allow to be measured (given that Big Data often involves large data sets, traditional sampling techniques could be used to acquire more meaningful subsets to work with).
- NSIs are not well equipped to conduct the primary processing and aggregation of Big Data, so outsourcing and partnerships with data suppliers requires further discussion. But this calls into question the future value of official statistics. One of the advantages of statistical organisations is their capacity to work with a broad range of data sources and statistical outputs. This places NSIs in a privileged position to provide validations, aggregations and to combine information from different data sources. In the future, Big Data will require a paradigm shift where analysis, rather than collection (or first-stage processing of micro data), is the primary mode of activity.
- If these changes are to be implemented, statisticians will have to convince users and suppliers of the value of these data sources.
- Managing discontinuity and change is a recurring concern expressed by statisticians. We are moving to an era where long time series are something of the past. Society is not as stable as it once was. The focus should be on producing timely statistics to measure contemporary phenomena.
- Statisticians need to be aware that users' behaviour can change and that this can create continuity issues. If users were to hide their digital trail as a result of privacy concerns, this would diminish the value of certain data sources.
- At present, NSIs have been opportunistic in their work with Big Data**.** What counts has been largely defined in terms of access, circumstance or legislation. Statistical organisations need to move away from this approach to develop more robust data strategies (e.g. incorporating secondary data sources into an overall data strategy).

10. These issues can be addressed through ongoing collaborations.

Collaboratory 3: Big Data and Urban Waste Management

Overview

Metrics are integral to urban waste management. While measuring and counting have historically occupied a central place in government practices, Big Data introduces a new kind of political arithmetic for public bodies such as the Greater Manchester Waste Disposal Authority (GMWDA). Concerns with smart cities, data integration, responsive systems, transparency and accountability are some of the many ways in which such data is helping to reimagine and realise contemporary urban government. Big Data is of increasing importance to contemporary urban transformation, providing the potential to produce efficiency gains, cost savings, improved public services, more robust policy outcomes and environmentally sustainable modes of urban living. At the same time, working with Big Data introduces a series of challenges regarding accessibility, representation, reliability, data quality, integration and implementation that make counting and measuring increasingly vulnerable and unstable. These are some of the risks and vulnerabilities that this collaboratory aimed to interrogate.

The collaboratory brought together social scientists and waste practitioners to explore the role and implications of Big Data for waste management by attending to the question: *How might big data enable and/or constrain the particular relationships involved in waste management: between public and private sector service providers, between levels of government, between service providers and customers, between public bodies and citizens?*

The collaboratory was structured around three areas of debate with roundtable presentations and discussions structured around a number of questions.

1) Data and Waste Management was structured around the following questions:

- How does Big Data differ from other types of data?
- Are new measures produced?
- How does data differ from information?
- How does Big Data do counting and measuring differently to statistical or administrative data?
- What is measured and what is valued?

2) Ethics and Openness of Data

- What are the possibilities and challenges of working with Big Data in urban waste management?
- Is Big Data open data?

- What are the possibilities and challenges of public-private partnerships for data management?

3) Policy and Behaviour Change.

- How can Big Data be used to shape policy decisions and respond to future challenges in waste management?
- Does it allow a different relation to the public?

These issues were canvassed in relation to the existing data practices of the Greater Manchester Waste Disposal Authority (GMWDA), in addition to waste practitioners' aspirations and expectations about how Big Data could contribute to the WDA's ambition of 'zero waste' to landfill by motivating behaviour change, altering consumption habits and encouraging recycling.

Participants included the Socialising Big Data project team, and social scientists and academics from Durham University and the University of Manchester, in addition to waste management practitioners from the Greater Manchester Waste Disposal Authority and local authority representatives from from Stockport Council, Bolton Council and Birmingham City Council. The collaboratory was organised by Prof. Penny Harvey, Dr Yannis Kallianos and Dr Camilla Lewis and held between 30 April - 1 May 2014 at the Manchester Cathedral Centre, Manchester, UK.

Speakers included:

- **Neil Swannick**, Councillor, Manchester City Council
- **Mark Muldoon**, University of Manchester
- **Celia Lury**, University of Warwick
- **Steve Rose**, Head of Strategic Research, Birmingham City Council
- **Nicky Gregson**, Durham University
- **Joanna Hayduk**, Greater Manchester Waste Disposal Authority
- **Peter Davies**, Research Officer, Greater Manchester Waste Disposal Authority
- **Catherine Alexander**, Durham University
- **Mark Newall**, Director of Resources and Strategy, Greater Manchester Waste Disposal Authority

Some background readings on Waste Management were circulated in advance:

- Gee, S. & Uyarra, E. (2014). Recycling and Waste Management Contract. Available at: http://www.gmwda.gov.uk/recycling-and-waste-management-contract
- Krenchel, M. & Madsbjerg, C. (2014). Your Big Data is Useless if You Don't Bring it into the Real World. *Wired*.: Available at: http://wrd.cm/1gS3oC8.
- Uprichard, E. (2013). Big Data, Little Questions. *Discover Society*, 1. Available at: http://www.discoversociety.org/2013/10/01/focus-big-data-little-questions/.

Summary of Presentations

Neil Swannick: Big Data – A Waste Perspective

Councillor Neil Swannick, from the Greater Manchester Waste Disposal Authority (WDA), discussed the future challenges facing urban waste management in Manchester, and how Big Data could data be incorporated into the WDA's ambition to reduce waste and encourage recycling. Measuring the content of bins is central to the WDA's approach to waste management because metrics based on waste compositional analysis determine how waste can be disposed of and recycled. Data recorded in 2011, for example, revealed that 74% of what is disposed in bins is potentially recyclable. This information can then be used to determine what type of technology should be deployed to facilitate recycling (e.g. the current 4 bin recycling system). Waste management is a dynamic process. It depends not only on recycling but also on consumption habits and, accordingly, waste management practitioners must target these areas in their entirety. The WDA have used sampling measures to analyse seasonal variations in collection rates, interpreting these results in relation to socioeconomic factors, public attitudes and behaviours. Results demonstrate variance among communities in terms of consumption, waste and recycling practices. Data based on performance indicators are then linked to information on waste composition.

There is a pressing need to discover new ways to change attitudes, policies and behaviours around waste in order to maximise resources because current waste practices are not sustainable. Such an approach requires a move from existing data practices, which involves developing new technical skills, as well as a cultural shift in how the population view and respond to waste. Big Data could be incorporated into the WDA's strategy to advance waste management practices. Once collected, such data could be compared among and within municipalities at both a local and international level. This would result in efficiency gains and cost savings that would then translate into improved services and sustainability outcomes. Big Data technologies could also be used to modify behaviour and to meet waste targets. By using digital devices and sensors to identify those individuals who are not recycling in Greater Manchester, strategies could be established to alter people's behaviour and motivate them to become more responsible citizens.

Despite the promise and potential of Big Data, adopting these technologies requires overcoming a series of challenges. In addition to technical and infrastructural issues, practitioners remain uncertain about how data could be used to change behaviour. While the population appears to support recycling, issues arise when attempting to translate beliefs into action. There is a general mistrust and ambivalence toward what are perceived to be intrusive technologies, such as putting incinerators in neighbourhoods or microchips into bins. In order to incorporate these technologies into waste management practices, the WDA will need to

establish public trust and overcome political opposition and ethical issues regarding privacy and surveillance.

Roundtable 1: Data and Waste Management

Mark Muldoon

Mark Muldoon, from the University of Manchester, discussed some of the opportunities and challenges of working with Big Data in terms of what can be measured and valued. He noted that while much Big Data is based on visualisation, these techniques are not always useful. For example, maps of small social networks can prompt interesting questions, but even moderately large data sets can produce indistinguishable results. When visualisation is effective, it is often because of some sort of emergent behaviour in which many identical (or similar things) correspond to produce novel phenomena not readily predicted from the properties of the parts (e.g. weather maps).

Big Data raises issues of data quality. This is especially due to the social network of corporate governance, where individuals have to rely on commercial services to clean and process the data, which can lead to sampling issues. Missing metrics is another issue associated with Big Data. While Big Data can be used as a predictive tool in an effort to deal with uncertainty, incomplete samples measured by imperfect instruments can cause problems for those working with such data. Minor uncertainties in the data, whether due to sampling problems or measurement errors, can produce major differences. Predictive computations of this kind are a second source of “bigness” in Big Data: modellers sometimes use data to create larger data sets.

Big Data software and analysis tools could open up new possibilities, many of which are unknown at this point in time. A case in point was a recent contest by the web-analytics firm, Kaggle, which revealed that because the volume of data and the requisite computational tools are much larger than the kinds of things that academics previously analysed, they constitute a new set of “tools” whose uses are not yet apparent, even to their developers. Tools have a way of finding new uses for themselves. Once a tool performs a task with ease, new approaches and views of what is feasible present themselves. In a similar way, much of the potential of Big Data is probably not yet realised.

Celia Lury

Celia Lury, Director of the Centre for Interdisciplinary Methodologies at the University of Warwick, presented a provocation based on the views of Sandy Pentland, a data scientist from MIT’s Media Lab, to stimulate discussion on some of the ways that Big Data is beginning to transform social scientific research. As an advocate of Big Data, Pentland highlights the possibilities associated with Big Data. He believes that data science is equivalent to social

physics with the capacity to revolutionise society. It is important to remember that what is emerging is a field in progress. At present, there is much uncertainty about what data science is or could be. The availability of Big Data software and processing tools is beginning to transform what we think data can do. In order to be useful, however, Big Data needs to move beyond methods that establish statistical significance. Generally, the volume of such data is so great that any question you ask of it will generate a statistically significant answer. This implies that the scientific method, as traditionally used, is no longer effective because the size of Big Data sets will almost always produce significant results, which can also lead to false correlations.

In order to overcome these challenges, Big Data requires human understanding of the connection between data points. This involves not merely contrasting cause and correlation, but steering a midcourse between them. The emphasis would move away from causality, which is bound to particular understandings of statistical significance, to a dialogue between our human intuition and Big Data statistics. What is required is new ways to test connections in the real world that move beyond laboratory experiments, collection and computational analysis to include real-life experiments. With social phenomena comprised of millions of small transactions between individuals, averages are no longer adequate as a way of representing society. Instead, individual transactions form the unit of analysis and it is these patterns that will provide the basis of this new era of social physics. Such insights have important implications for social policy in terms of what ought to be measured and valued.

Big Data raises issues around ethics and privacy. Consent and anonymity might no longer be sufficient or appropriate because these tend to remove the participant from the study, thereby, distancing the subject from the research. A way to overcome this issue is to think of Big Data experiments as participatory, so that participation is an active rather than a passive process.

Big Data focuses on computational techniques, but these must be understood in relation to a broader assemblage of research methods. These could be based, for example, on experiments in urban laboratories and iterative testing of correlations, which would provide a different way to establish the significance of correlations in which you imagine and experimentally test what could be significant rather than relying on statistical significance alone. Urban laboratories could involve actors in the coproduction of knowledge who participate in knowledge exchange, fostering an orientation to change built on a contingency of uncertainty. Working with Big Data in this way raises issues of scaling, generalisation, marginalisation and pathologisation, but it is an emerging mode of research. Simulation and visualisation could also be used not merely to illustrate a finding, but as part of the scientific method to identify significant relations. From this perspective, problems do not need to be hypothesised in advance. Instead, problems and significance could be produced simultaneously if visualisation is used as part of the process to establish significance.

Roundtable 2: Ethics and Openness of Data

Steve Rose

Steve Rose, Head of Strategic Research at Birmingham City Council (BCC), discussed ethics and open data in relation to the kind of data collected from fleet and waste services. Case studies were employed to demonstrate the innovative ways that the City Council has incorporated data into their practices to illustrate the potential of working with Big Data. BCC use the GPS data that tracks and records fleet vehicle movements to remodel the refuse collection rounds. This minimises distance travelled but also provides digital routes to start to link to customer relationship management requirements such as assisted collections. BCC also conducted a trial where residents were awarded Nectar points in response to recycling participation. Rewarding residents for recycling proved successful in terms of behaviour change resulting in 9% extra tonnage and a 10% increase in collection rates. Such initiatives suggest that customers and their data could be at the forefront of service thinking.

These case studies reveal the potential practical applications that could be realised from waste data and the role of open data in achieving this. There are opportunities to use Big Data technologies to develop incentives based on loyalty and reward. Big data could be used to improve council services, especially through data linkage and integration. Despite these potentials, it is difficult to translate trials into action and to change existing work practices. In order to be actualised, councils will need to develop new technical skills and partnerships (public-private). They will also need to overcome a series of ethical challenges regarding trust and accountability in order to encourage users to exchange politically sensitive information and to volunteer in the coproduction of such data.

Nicky Gregson

Nicky Gregson, from Durham University, explored the possibilities of Big Data in waste management, focusing on the potential and limitations of smart bin technologies and their association with the Internet of things. There was particular emphasis on the differences that certain devices and sensors could make to waste data and municipal waste relations. In the UK, the debate on data technologies is generally framed in relation to bin collection and understood in terms of track and trace, and chip devices. The problem is that the data generated by these devices only calculates specific metrics: weight (tonnage) and household (non)compliance. Such an approach provides a limited way of thinking about waste generation. These devices may be cost effective, but are relatively ineffective because they use the same metrics and subsequently as a technology are unlikely to alter behaviour, individually or collectively.

Big Data technologies, by contrast, such as smart bins (sensor devices designed to produce certain metrics) collect real-time measurement data. Data is measured not just in terms of

quantity, but also the *quality* of the content of bins (e.g. fill levels and temperature of materials). Technologies of this kind move data away from a reactive monitoring of a bin to the possibilities of waste data being anticipatory and predictive. This enables new possibilities of modeling, simulation, and forecasting waste data at the level of the individual household, which could then be amalgamated up to the level of the municipality. An example of such initiatives was conducted by the Scandinavian company, Enevo, which used smart technologies to produce efficiency savings in public services. Sending trucks on near-time modeled routes, instead of predetermined routes, resulted in a 30-40% reduction on logistics costs. These technologies have the capacity to change the relationship between the household and the council, leading to more personalised, cost effective services. To benefit from these opportunities, however, requires a high degree of public trust, which is arguably absent in the UK. At the same time, working with Big Data technologies raise associated issues of public fear; resistance; data integration and work cultures.

In sum, the shift towards Big Data signals a policy move from measuring waste in terms of tonnages of materials diverted from landfill for recycling to secondary resources circulating within a circular economy. These real-time (or near time) metrics will assume increasing significance in the future for waste practitioners. From this perspective, it is important to think about waste management as resource harvesting. Such an approach has implications for data collection with an emphasis on the *quality* of content as opposed to standard measurements of quantity and weight (tonnage).

Joanna Hayduk: Using Data to Inform Decisions

Joanna Hayduk, from the Greater Manchester Waste Disposal Authority, discussed how local authorities use data to inform decisions. At present, LAs use a variety of data sources. This includes sociodemographic, operational and cost data, as well as recycling service information, made available at a local and national level. Examples of how LAs use data were then discussed in relation to two case studies: WasteDataFlow and WRAP's Local Authority Waste and Recycling Information Portal.

Big Data introduces new possibilities and has the potential to lead to efficiency and cost gains. At the same time, working with such data presents a series of challenges regarding availability; relevance; representation; skill shortages and the inability to collate and process such data. While data linkage could provide new insights for waste practitioners, public-private partnerships tend to be hindered by ethical issues of privacy and confidentiality, particularly given the different objectives of these groups.

Roundtable 3: Policy and Behaviour Change

Peter Davies: Socialising Waste Data: The GM Experience

Peter Davies, from Greater Manchester Waste Disposal Authority (GMWDA), discussed the WDA's experience of using waste data to encourage recycling and behavior change. Greater Manchester Waste Disposal Authority consists of nine District Councils, which have been brought together under a 25-year Recycling and Waste Management Contract. The majority of waste is delivered to facilities via a four-stream kerbside sort system, or taken directly to household waste recycling centres. Weigh-bridge data provide accurate figures to assess the performance (e.g. landfill diversion, recycling rate and the contract), but is ineffective in helping practitioners develop new strategies to encourage waste prevention and recycling.

In 2011 a waste compositional analysis was undertaken. This study provided a better understanding of what materials were not being captured and the type of households and demographics that were not recycling. When combined with national data, this information was used to formulate the basis of the LIFE+ project, which aimed to target an increase recycling practices among specific groups (e.g. rental properties).

The nature and variability of waste has raised significant challenges for the WDA, particularly in terms of identifying areas of low performance and assessing the impact of campaigns across the population. These issues need to be addressed in order for waste practitioners to inform policy and behaviour change. Public and private partnerships could help to overcome this issue. The academic community could also provide valuable input by exploring new ways that data can be used to support behavioural change.

Catherine Alexander

Catherine Alexander, from Durham University, discussed some of the opportunities and challenges of using Big Data in the context of waste management, sustainability and the environment. These issues were canvassed in relation to a project that explored recycling habits and third sector recycling companies in the London Borough of Southwark. Together with colleagues from the Environmental Engineering Department at the University of Surrey working on the project called LARA (Local Area Resource Analysis), the project sought to understand why areas featuring dense housing and high-rise estates had particularly low recycling rates. Attempts to collaborate with other researchers working on similar studies raised issues of comparability, data linkage and integration given the different practices of collecting and analysing the data, particularly when expenditure was used as a proxy for consumption to predict future household waste, as was the case with LARA.

In addition to these methodological challenges, several insights arose when comparing methods to predict waste by looking at household metabolisms. While the majority of studies that examine material flows take place in households, they all use slightly different methods based upon different assumptions and metrics (e.g. using expenditure as a proxy for consumption instead of metrics based on weight). In effect, this means that although there is much data to reference, there is no cumulative data object or "thing" that researchers can track, compare and develop. Other challenges that arose concerned missing metrics, incomplete or unreliable waste statistics, particularly in deprived inner city areas. This raised associated methodological issues regarding classification and representation. Census data proved particularly unreliable as it had a low rate response rate (52%). These issues were confounded as census classifications did not always correspond to household practices (e.g. sharing income, expenditure and differences between short-term and long-term rentals), just as expenditure data did not adequately represent household economies (e.g. childcare and income). Data of this kind resulted in a smoothing effect that failed to represent individual variation between households or reflect the rapid turnover of people of certain populations (due to relocation or death, for example). As a consequence figures, models and assumptions had to be re-evaluated. Moreover, the assumption that you could predict the timing of waste arising by the end of the functional life of something was problematic because other factors come into play (e.g. high churn rates). In revealing the tendency to infer a general pattern from the individual without thinking about individual variations, these studies highlight the need to critically assess the data upon which urban waste management is based.

Mark Newall

Mark Newall, from the GMWDA, discussed how Big Data can be used to shape policy decisions and respond to future challenges. Waste management needs to be understood in the context of sustainable development and limited resources. The population is projected to rise from c7 billion in 2014 to c9 billion in 2050 and as developing countries continue to grow the demand for raw materials is increasing and C02 emissions, particularly in Asia Pacific countries, are rising rapidly. This requires a fundamental change business practices and how the world consumes that requires the public, government and commercial sectors to work closely together.

Behaviour change is the key challenge currently facing waste practitioners. Consumers need to be the driving force for sustainable goods and begin to see waste as a resource. Education, high-quality data and consumer feedback (via campaigns and social media, for example) is essential if behaviour change is to take place. In order to be effective, these must occur within a robust policy framework. To do so requires data linkage on waste data, the environment and climate change and sustainable consumption and production by engaging with the policymakers, businesses and consumers. Only then can new practices around waste emerge that will result in substantive effects.

Summary of Key Themes Arising at Each Collaboratory

Following each of the collaboratories, the project team reflected on the various presentations and the conversations and discussions that ensued. For each collaboratory we analysed and organised the outcomes in different ways. For Genomics, in relation to key challenges; National Statistics in relation to key questions; and Waste Management in relation to key topics. These are of course not exhaustive – numerous issues were raised. Here we sought to highlight what we considered to be some of the most prominent and repeated. Like the conduct of the collaboratories we trialled different ways of doing this as is reflected below.

Collaboratory 1: Genomics

Indeterminacy, Uncertainty & Errors

- There was an acute awareness of having to work with errors and uncertainty. One manifestation of this is the challenge of distinguishing between signal and noise; whether a given sequence variation is biological (i.e. real), or due to experimental artifact or error? Other points of indeterminacy come from the challenges of mapping a given sequence onto the reference genome. This challenge is compounded by the repetitive nature of parts of the genome combined with very high focus close-ups of the genome (short reads), making it like doing a jigsaw puzzle that has lots of blue sky. Indeterminacy is also generated by the changing nature of the reference genome. A further area of indeterminacy is interpretation – even if the sequence is correct and correctly mapped – what does it mean?
- Genomics practitioners are very aware of the *performativity of their methods*. They live and work in knowledge of the reality that different instruments and different software analysis tools find different data in the same sample. For example, when instrument vendors change their quality control software, then the sequence changes!
- This means having an awareness that having more sequence data produced by 'next generation sequencers' (NGS) wills not necessarily reduce uncertainty. On the contrary, it can increase it. For rare diseases, e.g., as the number and intensity of NGS studies increases, the less certain the meaning of a single rare variant becomes. Similarly, in terms of sequence coverage, there is a need to find a 'sweet spot' in between too much and too little coverage. The same applies to software tools. Instead of triangulation, the more analytical tools that are used, the more the uncertainty can proliferate.
- Is it possible to step outside of NGS to make a reality check? Is there a Gold Standard for comparison, for example, is Sanger sequencing used to determine discrepancies in NGS data? Not necessarily. Again, practitioners at the workshop live with an understanding of the performativity of *all methods*, not just NGS. In other words, there is no 'sky hook' – no privileged position – from which to get unmediated access to the real. Yet work continues despite the difficulty of knowing what is real and what is artefactual or error. Radical indeterminacy is not a deterrent to progress. Moreover, 'this is nothing new'.
- When it comes to interpretation and prediction, e.g., severity of a disease, NGS data has to be compared with and related to other forms of data to get an 'orthogonal' check on what it means. Selecting what to compare a given set of results which requires professional

judgment and expertise, including an assessment of the reputation of other scientists and their laboratories, which may determine whether to take their results into account or not.
- However, comparison between different data types is not straightforward. Integration is problematic. There is awareness that comparison does not necessarily lead to convergence since different assays on the same sample may not readily align with a singular underlying reality. Consistency between different kinds of data is itself a quality that has to be assessed and measured and produced rather than a given, as was illustrated in the case of multiple (omic) studies of stem cells.

Diversity and Difference in NGS Producers, Uses & Applications

- NGS data are produced for a wide variety of uses, communities and applications. These different real and imagined downstream deployments of NGS dataset and their interpretations are apparent in how they are evaluated, packaged and stored.
- One area of difference is the validity of result and the rigor of their testing. At one end of the spectrum are clinical applications. For this, tolerance of uncertainty is very low because lives may be at stake. 'Are the results and analysis reliable enough to be the basis of counseling for therapeutic termination?' Further down the spectrum is the realm of scientific knowledge production. Here, whilst uncertainty is to some degree minimized, the consequences of acting on it are not as critical; indeed, individual scientists and laboratories may be tempted to exploit lack of certainty by selecting modes of analysis that yield the most 'interesting' or publishable findings rather than those that are the most rigorous. Another category is large-scale community projects, such as the 1KGP and the International Wheat Genome Sequencing Consortium, that have relatively more resources and which check all stages of the process in order to deliver a quality product that is expected to have widespread future uses.
- Another variation that reflects community differences is the degree to which NGS datasets are neatly packaged and stored, and when and whether they are made publicly available. Large-international community projects, notably 1KGP, have an 'aggressive' metadata standard and the resource, including at the key repository (SRA), to ensure they are fully documented, and a commitment to public deposition. Individual labs may only be motivated to make their datasets publicly available in order to publish, and then have only the time and resources to provide minimal metadata. Then there are the myriad commercial NGS datasets that are privately held (but whose meaning and value can only be realized through comparison with public datasets).
- Differences among uses and applications of NGS were also apparent in relation to notions of trust regarding stewardship of datasets. Although Amazon Web Services currently stores a number of publicly funded NGS datasets, including 1KGP, can it be trusted with clinical datasets, e.g. from the 100,000 Genome Project?

Trust & Commercial Actors in the Ecosystem

- Commercial actors are integral to the NGS ecosystem, e.g., as instrument and informatics vendors, data service providers, and data storage providers.
- It was accepted as a matter of course that the source of data would be instruments that are commercially produced and sold. This is something that genomic scientists – and scientists

more generally – are accustomed to. Perhaps because of this, the centrality of instruments did not seem to engender distrust of the instrument providers as threats to the supply of, or access to, NGS data, or control over its analysis (cf the situation with respect to national statistics). Commercial instruments and their providers (and commercial facilities that operate them) are a familiar and accepted part of the NGS ecosystem. Moreover, NGS practitioners are fully in control of the samples (again compares with national statistics). Historically they have been originators of the analytic algorithms (cf Google Analytics and national statistics). However, some of quality control algorithms are built into the instruments, and when these change so does the data!

- This trust of the NGS ecosystem was in contrast to the position of AWS (Amazon Web Services) whose participation in the ecosystem as provider of data storage and analysis in the cloud was not universally trusted - although, as one participant questioned, why trust Illumina and not AWS? Coming from a different context and a newcomer to academic and clinical research, AWS was felt by some participants to lack 'street credibility', with much to prove as a credible and trustworthy custodian of clinical data.

Middle Phase NGS

- There was a keen sense of periodicity from some of the speakers; that NGS is now in a second, or middle phase – that there was an earlier period of genomics ('Noah's Ark') that was different to the present, and a future phase that that is going to be different yet again. This was expressed in various ways using a variety of metrics.
- Compared to the past, the present is marked by increasing numbers of genomes being sequenced, and increasing volume of sequence data. We are in a different phase from the 'Noah's ark' period, when the big genome period studied just one of everything to using NGS for testing hypotheses and experimentation. However, this is matched by falling costs of sequencing and falling costs of data storage. Therefore, for NGS, according to one speaker, the challenges of Volume and Velocity are a thing of the past. The challenge today is Variety.
- Even the speaker who pointed out that the dramatic fall in sequencing costs was leveling out, agreed that the cost bottleneck is no longer sequencing. The cost bottleneck today is upstream and downstream of sequencing, and looking to the future, this is liable to be the case even more so. Another third kind of uncertainty was in relation to costs, in particular the costs of analytics. With the broadening of the applications of NGS from community sequencing activities to testing hypotheses through experimentation, the costs of analysis become more unknowable in advance as they attempt to decipher the complexities of living systems.
- Whereas in the past, 10-15 years ago, the human eye would pass over all of the data as teams of people analysed it in depth, nowadays the data generated are never fully analysed.
- In terms of an 'innovation timeline', the past was the 'invention' phase; the present is the 'development' phase, and the future is the 'productisation' phase.

Metrics

Participants from different parts of the NGS ecosystem attend to different risks and uncertainties; they have different priorities; and they relate what they do to different metrics. These are apparent in the three 'ideal' types that were initially identified by the project team prior to the collaboratory and then 'mapped' onto the different organisations and practices.

- These are the practitioners involved with creating sequencing data and making it fit for others to use. These practitioners are focused on the integrity of the sequence – that it is accurate, in the right order, and that it is what it says it is, i.e. that the chain of custody is secure. They are providing and curating and resource for others to use – either individual clients or communities.
- BBSRC Genome Analysis Centre (TGAC) – turning samples into sequence - generating sequence as a service – samples come into TGAC, sequences go out. In between, the sample goes through instruments. But the data that come out of the NGS instruments are in bits. All these bits have to be reassembled in as good as order as possible within the available analytical and financial limits. Can they bring it all back together again? The speaker from TGAC relates to metrics about species and their genomes (how big? how many repeats?) and how best to package it for the instruments; and to metrics about the sequence data; and metrics about analytics – N50 and K-mers.
- Centre for Genomic Research. Leads a lab that uses NGS data and instruments to address biological questions, to test hypotheses. Is aware of instrument metrics – cost of instruments, costs of sequencing, speed, read length. Is also aware of costs of sample preparation and, in particular, the costs of analytics. These are the new bottlenecks. The costs of analytics, in particular, is becoming increasingly unknown and unknowable in advance. This source of uncertainty is increasing as NGS is used for purposes other than sequencing, e.g. for doing biological experiments. Indeed, this source of uncertainty is a reflection of the unknown complexities of biological systems compared to the relative simplicity of the molecule DNA.
- These are the practitioners who use and apply data in novel ways. Further downstream, arguably they deal with more unknowns and unknowables than classes A and C because they are integrating data with other data and with biological, social and economic systems.
- Rare Diseases - using NGS data to try to identify what causes rare diseases for knowledge that can be used in clinical practice. On the hunt for genomic variation. But aware that variation might not be real – it could be error. So uses very, very high stringencies that miss things in the interests of only finding what's real. And even if it is real, it might not be biologically meaningful. Because of NGS as a unified way of studying them, rare diseases are becoming common. By lumping rare diseases together as having a common cause – a genetic basis detectable in the DNA – they are becoming a bigger number – even if they are all different from each other. Balancing uncertainty and risk are key qualities to grapple with. As a clinician, the key question is 'Am I certain enough – are the data certain enough – to sanction termination?
- Public Health Genomics – translating NGS into NHS. Here the metrics are expressed in terms of costs and benefits. But the limits in time and space are hard to draw – whose costs?

Whose benefits? When? And in relation to what? More questions than answers. Raises questions rather than relating to metrics because the metrics are too hard to define as you move further downstream to application, particularly clinical application.

- Eagle Genomics. Eagle uses publicly available NGS data to add value to / make sense of private / commercial NGS data sets. Is not the volume or velocity of NGS data that is the challenge; it is its heterogeneity.
- These are the practitioners who are concerned with the bulk handling and warehousing of data, with making it available and discoverable for community use and providing the means to access and work with it.
- EMBL – EBI Resequencing Informatics. Processes the datasets from large scale community projects. Is focused on projects and their resultant datasets. Is a quality-control operation. Wants to scrutinise the quality of the stuff that is going to be put in the warehouse before it is distributed. What is their quantity – how many genomes, how many Nbytes? What is their quality – do they make sense biologically, what virtual analyses can we do to test this? Are they what they say they are - their identity? What is their consistency with other kinds of data?
- European Nucleotide Archive – responsible to submitters and users of the SRA / ENA. Is concerned with discoverability and accessibility of data. Tension between obligation to receive and archive everything and dealing with the challenges this presents in terms of quantity and quality – and how this impinges on discoverability. Focus is on the qualities of submissions – this is the key unit of analysis? Looks at doubling rates. Metrics that do **not** feature here are number of repeats, speed and cost of sequencing.
- Amazon Web Services – provides a space where people – customers – can store and analyse their data. Is about its 'instances' whose qualities are computational. The data are *containerized*. Black boxes. – Nothing biological about the metrics that this speaker uses. The black boxes could contain anything? Metrics are size and cost and speed and flexibility. Size is not a problem. Nor is speed. Nor is cost if you buy your space in the cloud far enough in advance.

Collaboratory 2: Official Statistics

One consensus emerging from the workshop was that this is an uncertain time and to what extent Big Data (BD) sources will be used to replace, supplement or verify official statistics is yet to be seen. All scenarios have potential risks as they raise fundamental questions about the substance and meaning of what is measured, captured and counted and in turn what comes to matter and count as official statistics. Some of those risks are perhaps a result of often-implicit normative assumptions about BD sources and what they are measuring. By unpacking some implicit assumptions those risks can perhaps be identified and debated, which is something that ideally should happen at this deliberative and experimental stage.

Generally the risks pertain to what we could call the qualities of data, measurement and what is valued, which tend to get obscured perhaps because of the focus on questions of data volume and analytics. Following the workshop three themes posed as questions arising from the presentations and discussions were identified and shared with the participating statisticians:

1. What is measured is what is valued?
2. What people do but not what they say?
3. Privacy and confidentiality but wither ethics and consent?

Through written responses and conversations at subsequent meetings and events it was agreed that these themes covered the meta-level issues and controversies raised at the collaboratory. Specific feedback on was annotated below and gives a flavour of the iterative process adopted in this collaboratory (*noted in italics; bullet points indicate comments from different people*). Additional themes were also added that address other issues and dilemmas arising from these questions:

4. Improving processes rather than producing outputs?
5. Not just methods but a change in organisational cultures and thinking?
    a. Data science is not a person but a distributed set of skills?
    b. From NSI-lead to collectivised processes of innovation?
6. A new statistical leadership role for NSIs?

1. What is measured is what is valued?

BD sources change what we measure and then value. What is counted and counts is driven by what data is available and not only or necessarily by the questions posed and classifications imposed by statisticians or others. Does this mean that what is measured is more reactively than actively defined? If the answer is yes, then this potentially has policy implications. One can imagine, for example, that the issues, concerns and priorities of governing could be influenced and driven by the kinds of data organised and configured by others, most significantly,

commercial operators. Do 'others' such as commercial operators then become more important players in determining what comes to be measured and then valued?

- *The private sector has moved ahead of us and we are behind. Industry operators such as Google are investing more in Big Data than NSIs and attracting the best people from around the world.*
- *At the same time, to work with BD requires establishing new partnerships outside of statistics and especially working with data owners, researchers and different user groups. In other words, it is not possible to work separately from and competitively with the private sector but instead find new ways of working in partnership.*

Do those issues and phenomena that are most easily and readily compiled digitally become key drivers of what then is measured and valued?

- *These inquiries seem to question the role of Official Statistics. I would indeed reformulate them in a different way, namely: Official Statistics is what policy makers need, because it provides the guarantees necessary to make good and responsible decisions. However, BD can further improve Official Statistics in terms of timeliness and new phenomena that can be studied. Here, BD plays the major role with respect to Official Statistics, and not giving over to the possibility that other players could replace Official Statistics.*
- *Companies are being driven by data push and we can learn from them: instead of searching for an answer in a data set (data pull), start with data and see what questions you might be able to ask (data push).*
- *Not only does data drive what is measured but so too does it depend on the available technologies.*

A related question is that BD sources are often cited as being less resource-intensive and less costly and that these are key drivers for taking up these sources especially at a time of budget cuts. At the same time, the 'actual' costs of using BD sources at present and into the future are unknown given that industry providers own much of this data and rigorous use of these sources for official statistics may well require purchasing data. So cost is both a driver of what is measured and also a source of economic uncertainty and vulnerability.

These questions are connected to the 'new' and novel in what BD sources measure and count which needs to be specified rather than assumed. Given the variety of sources—from search engine queries to social media messages and mobile phone usage—what is being measured and its relation to previous forms of measurement are diverse. Thus while BD is usually reduced to a notion of some kind of generic 'data' we know that not all data is the same. Towards thinking about this, here are some suggested ways of thinking about what are 'old' and 'new' measures:

1. *New phenomena and new measures?* Information and communication technologies generate new relations, practices and phenomena that are now also the object of measurement such as ICT usage, which for example has become a measure of an 'information society.' Here both social relations (how people interact and communicate for example) as well as the measurement of society are simultaneously changing and being done together. A good example of this is the medium-specific elements of a platform; for Twitter this includes retweets and hashtags that organise associations and social networks in new and novel ways and also open up those networks to data analysis.
2. *One phenomenon for another?* Some forms of measurement are becoming more relevant than others in part because of the characteristics of a source and the volume of data. A good example of this is Twitter, where sentiment analysis could potentially replace opinion surveys. What is the difference between sentiments and opinions and does this matter? Perhaps (as suggested in the second theme below) sentiment is not a quality that is intentionally conveyed but read-off of messages and presumed. There is a long history behind the acceptance of opinion research as a 'real existing' social phenomenon that can be measured; can the same be said of sentiment?
3. *Old phenomena and new measures*? Some BD concern phenomena that could not be practically measured very well or at all in the past such as the spread of rumours or the purchasing intentions of consumers or the movement and travel patterns of tourists. Here, what previously was not is now rendered measurable. Do such phenomena become 'elevated' and promoted as relevant and those that remain 'old' and unmeasurable (e.g., rationalities) remain obscure?
4. *New phenomena and old measures?* BD sources, in part because they are unstructured, contain what is sometimes referred to as noise or excessive content that needs to be cleaned or reduced. Implicit in this evaluation is that what is of value is that which is recognisable according to existing frames and understandings. Following from (1) above, perhaps, the content of BD sources reflects an incongruity between existing categories and classifications and the kinds of phenomena generated or registered by a platform.
    - *Instead of talking about an incongruity between BD and the Official Statistics framework (design metadata, classification, etc.), I would stress the fact that in Official Statistics you first 'design and then 'collect', with BD, instead, you first 'collect' and then 'design'. So, in a sense, a completely different 'process.'*

      This incongruity is not a unique condition of BD sources; other methods generate data that fall 'outside' of imposed classification schemes and while sometimes excluded at other times these have changed classifications. This is evident in the assertion of new categories in a survey or census as in the case of the Jedi religion in the UK census.

      Instead of reproducing results compiled by methods such as surveys, is the issue then that BD measures phenomena that do not correspond with 'old' measures and so instead of comparing the challenge is to innovate new measures of society? Official statistics begin with 'perfect' classifications – is it now the other way around whereby the issue is how to create new classifications that fit with BD sources? For

example, the UN Global Pulse initiative analyses Twitter for signals of distress rather than attempting to make meaning of the messages. Distress is identified as an anomaly in a regularised pattern of sentiment internally constituted via the platform.

- *Users of Official Statistics are familiar with traditional and longstanding measures. Big Data sources introduce new kinds of measures such as signals and indicators of change that are also timelier (e.g., UN Global Pulse). Official Statistics, whether from admin data, surveys or censuses, cannot match this; at the same time, signals and indicators cannot provide the quality, consistency, depth and trust that Official Statistics can. It is thus perhaps a question of the complementarity between these two sources rather than a question of one or the other. One criterion of assessment, which is central to the purposes of Official Statistics, is to assess the extent to which a particular kind of BD is useful for decision-making purposes and policy evaluation.*
- *There is a need to challenge the assumption about what BD versus Official Statistics measure. Instead, why not start with what are the important questions and problems today and what data can answer these? Can phrase this as 'if Official Statistics can't answer questions that matter then they will not be relevant.' And this includes timeliness—some questions need faster answers.*
- *If you deal with these issues as output problems and evaluate them in relation to existing measures, then you will always end up back with your own source and method. That is, BD calls for not applying existing methods and adhering to the same tests of validity and quality.*
- *At the same time, Official Statistics can serve the role of providing a benchmark to test new sources against and that is potentially an important value.*

5. *From snapshots to societies in process*? Much is said about the timeliness of BD as a positive quality. At the same time instabilities, discontinuities and uncertainties about BD data sources over time are identified as a threat to time series types of analysis. Yet there is also recognition that BD is more immediate and this immediacy is one quality that is valued. What then are the implications for decision-making? Does decision-making become based more on what is happening now, on societies in-the-making rather than on snapshots of today/yesterday and change?
    - *Coming at this question from another vantage point is that this suggests stabilising statistics based on existing and approved methodologies. How can this be done in a world that is becoming more agile and flexible? It is not possible. New paradigms and methodologies are required.*

2. What people do but not what they say they do?

BD sources signal a move from data provided by respondents to that generated as a byproduct of what they do. That is, through methods such as surveys and interviews respondents can choose how and what to inform about themselves. The same does not hold for their digital traces and how those traces may 'speak' for them.

Ostensibly, one reason byproduct data is deemed valuable is because it reduces 'respondent burden', which in part is reflected in declining response rates. But, as raised a few times in the discussions, respondent data is not the same as byproduct data. The former is stated and measures a response and the latter is a register of behaviour. The latter more strongly apply to online BD sources but arguably could also hold for administrative sources, which are the byproduct of administrative transactions and not intended for statistical purposes. Here are some possible implications of these differences for further discussion.

1. *The respondent*. Much has been said about problems of respondent memory, the accuracy of self-reporting and the problems of non-response bias. That is, people are not always cooperative, reliable and responsive. But rather than resolving these issues do BD sources just shift and redistribute the 'problem' of the respondent? Users of online platforms or administrative clients are not necessarily more reliable, accurate and responsive. People can modify their behaviour and use of platforms, opt in and out of location services to avoid or prevent detection, and create several versions of themselves through multiple devices or profiles.
    - *The failure of Google Flu Trends to predict trends in flu cases is a good example. Over time it became less reliable because people's behaviour changed. Google called this 'feedback' - people reading about Google flu trends, googling about this and because of the autosuggest feature of the algorithm their googling contributed to the trend leading to an over-estimate*.

    Thus there are many complications surrounding the 'who' is the data about such that non-humans also become a consideration—is a data trace a measure of the behaviour of a device (car, mobile phone, computer, scanner, sensor), platform (social media, browser), algorithm (search engine, bot) or person (purchaser, browser, tweeter)?

    Finally, does it matter for policy whether data is intentionally provided via people reporting or being asked versus their behaviour being known via tracing and tracking?

2. *Causality and meaning making*. With the promotion of behaviour, 'deeper' questions such as why people do what they do and the meaning of what they do are potentially demoted. Some insight about this could perhaps be gleaned by explaining what seems like the (uncanny) correspondence between survey results on consumer confidence and Google trends or between a migration survey and Google trends. Is this correspondence a validation of survey results? Or is it a classic example of a spurious correlation? Are these two methods measuring the same phenomenon/ behaviour or is the correspondence a measure of a common trend between two different phenomena/behaviours? Is this an indication of the move away from a concern with causality and towards more pattern analysis?

3. *Representativeness of the data*. The questions of what and who is a BD source is measuring also extends to questions of representativeness. Much statistical work in the past has attended to this question to ensure that results are not biased towards particular groups in a society. Since socio-demographics are typically not attached to many BD sources, then what can be said about who is being measured and included? Rather than society, do BD sources measure users of platform such as Twitter and Google?
    - *This is less an issue when doing time series analyses for identifying change. Though this kind of analysis is self-referential - the population covered results in a pattern and when that pattern changes over time then this is a 'signal' that something new is happening within that population. However, what of populations not included (since people are self selecting on a platform) in the time series and which cannot be identified due to privacy issues? If they were, would the pattern be different? Over time is the same population being measured?*

3. Privacy and confidentiality but wither ethics and consent?

Historically and increasingly so, NSIs and their governments have made great efforts to ensure that the privacy and confidentiality of individual data are secured. A variety of techniques and rationalities have been used to do this such as anonymisation and protections against disclosure. Additionally, privacy is said to be secured when the object of concern is trends, patterns and aggregates rather than the details and specificities of individuals. Finally, practices of reusing data are said to be less intrusive and thus more respectful of privacy. These arguments implicitly assume that ethical concerns are principally about the identification of individuals. However, should ethical concerns also extend to questions of consent and intent?

- *There are also potential group privacy effects that cannot be resolved by the anonymisation of individuals. That is, even with individual privacy secured there are implications for the identification of groups who then can become the targets of policies.*
- *Transparent use is one possible response. That is, treating all data as personal and thus applying the same protections no matter what the source is another*.

These issues are also matters of concern to academic researchers who question the ethics of presuming that because data is generated in the 'public' space of the internet, the consent of subjects is not required. For official statistics this potentially becomes more complicated if digital traces are used for or influence public policy decision-making. Again the comparison to other methods is useful here. When respondents give an answer to a government survey or census questionnaire they are aware that this data will likely be used for public policy purposes. The same cannot be said for BD sources. The intent of a user's engagement with a social media platform is to communicate, be social or get information and not to serve the governmental (or

research) purposes of others. Is one implication that the active involvement of respondents in shaping or influencing the evidence-base of policy is undermined by the use of BD sources?

Because of the issues of privacy and ethics, might BD sources be leveraged more for measuring things—traffic, cars, crops, water, or prices—rather than people—tweets, messages, profiles or searches?

- *We do agree that we are in an experimental phase. However, given that BD sources can be very much different in terms of intrinsic characteristics and access modalities, even at this stage some typologies seem more promising in terms of possible usage on the short term. In particular mobile phone data and data deriving form 'traditional' transaction processing systems (e.g., scanner data) appears more promising to be used for OS purposes in the short term.*
- *This issue of measurement is taken up in the proposed UNECE classification system, which is mostly event based (rather than about separate individuals) with a move away from the 'unit' to the 'event', or from people to places (e.g., satellite images):*
    - *Human-sourced information (Social Networks)*
    - *Process-mediated data (Traditional Business Systems and Websites)*
    - *Machine-generated data (Automated Systems)*

The following additional questions arose in response to the above questions:

4. Improving processes rather than producing outputs?

In the face of the many questions such as those raised above, the potential of Big Data to inform statistical and other processes could perhaps be strategically the best focus of experimentation. That is, rather than testing BD as a source of Official Statistics, how can BD be used to monitor, evaluate and improve statistical processes and in this way not only be low risk but enable early experimentation and build capacity working with new forms of data? For example, how could smart energy meters be used to identify vacant properties and the optimal times that people are at home and available for enumerator visits and thereby improve response rates and decrease revisits?

With the increasing move to online censuses, government services and surveys, the potential of paradata (usage and behaviour data such as clicks, duration, pages read; also field operational data) is another potential area for experimentation. While still raising questions of ethics, if made transparent, paradata could be used to improve statistical data collection processes.

This aspect of BD, which could be called a process orientation, cuts across and addresses all of the three themes above. It changes the focus of each of the themes—whether about measures, respondents or ethics—to how BD can serve as a supplement to and improvement of Official Statistics.

5. Not just methods but a change in organisational cultures and thinking?

Many of the issues raised at events and meetings involve the repetition and sometimes over generalisation of issues. There is also much resistance expressed to change, of statisticians holding back and waiting to see what others do, etc. The initial reaction of statisticians to the new opportunities afforded by BD was previously and for some continues to be that of scepticism. They are generally risk adverse and come out of rigid institutional structures. One definition of BD offered that reflected this and which could be interpreted as a search for a way to retain authority was that of 'secondary data'. An alternative reaction was to re-brand statistics as 'Greater Statistics'.

In response to these observations and arising from a number of conversations, it was suggested that part of instilling a new culture involves the reiteration of statements as a way for a community to emerge around common problematisations and concepts. That is, existing and accepted methodologies such as censuses and surveys emerged over a long period of time and debate through various transnational forums and practices. At certain points in time what could be called a relatively stable 'technical settlement' was reached. That is perhaps one way to describe the current situation with respect to BD. However, there are many differences. In addition to the issues of ownership of data and reliance on the practices of others (e.g., industry or admin data owners) a different and experimental approach is required. As expressed under the first theme, BD does not call for the usual practice of 'designing then collecting' but 'collecting then designing'. The latter involves an experimental orientation, of working with data to test its qualities, uncertainties, capacities, patterns and so on and through which new methods, concepts and measures can be generated and debated and new technical settlements can perhaps be established. To do so requires organisational cultures that are agile and flexible, and able to engage in experiments, simulations and modelling. As the sandbox and innovation labs have shown, these require not only new computational techniques but new technical infrastructures.

a. Data science is not a person but a distributed set of skills?

Many comments note that the required skills to work with BD are distributed amongst a team – methodologists and IT people - rather than residing in any one person such as a 'data scientist'. 'Data science' is also described as a mindset rather than a person.

b. From NSI-lead to collectivised processes of innovation?

One model of innovation that is being experimented with involves statisticians from across NSIs working together with BD to test and model phenomena (e.g., UNECE sandbox project) as opposed to the more usual decentralised practice of NSI's developing new methods and then sharing with others as best practice. Though the latter is often mutually informed the former is a different site and practice of innovation. Such collaborations with various experts and

stakeholders are also a means of reducing investment costs and developing methodologies especially at a time of cutbacks and austerity.

6. A new statistical leadership role for NSIs?

BD is an opportunity for NSIs to show responsible statistical leadership through their adherence to the recently adopted UN Fundamental Principles of Statistics (impartiality, reliability, relevancy, profitability, confidentiality and transparency). This could include providing accreditation or certification on different data and measures.

Statistical organisations already have an advantage: their capacity to work with a range of data sources and statistical outputs. They already have long-tested skills validating, aggregating and combining data from different sources. This may be a future role where analysis rather than collection is the primary mode of activity.

Collaboratory 3: Urban Waste Management

Policy and behaviour change

In the last ten years, the Greater Manchester Waste Disposal Authority (GMWDA) has made great improvements in diverting waste from landfill (from 4% in 2003 to 55% in 2014) but the Authority is striving to improve on this target and to ensure that there is 75-90% diversion by 2019-2020. By this date they have also set themselves the ambitious target of an overall 42-50% recycling rate. In order to achieve these aims, they are testing a number of different types of public engagement and using a variety of communication campaigns to try to create behaviour change. To reach national and EU targets, the authority needs to maximise the quantity and quality of recyclable materials that they collect. This requires collaboration from the public to sort their rubbish more accurately. A Waste Composition Analysis conducted in 2011 found that 74% of the waste in residual bins could be recycled. An EU LIFE + project has been exploring the impact of a number of communication campaigns on 'low performing' areas. The project is looking at tonnage data and trying to match this data to metrics of participation of households and the performance of each authority. It has come up against many methodological challenges. Waste is inherently complex and behaviour patterns are difficult to detect in the current data which is available. Social scientists point out that obtaining accurate data on waste in urban areas is notoriously difficult, particularly in areas with dense and mobile populations. There are many connections which cannot be explained. For example, it was expected that low recycling rates would exist in areas with high deprivation but this was not uniformly found in the data. Also, in some areas after communication campaigns the tonnage of waste collected had gone down but the participation has gone up, or vice-versa.

In recent years recycling rates have started to level off. It is questioned whether residents have reached the limit of habitual change. Social science research suggests that recycling rates should be understood in relation to existing social practices and consideration of the infrastructural difficulties for recyclates storage and collections. The Authority argued that they must bring about 'behaviour change' among residents by encouraging education about resource efficiency and promoting wider 'cultural change' around sustainability. They are interested in whether Big Data could assist them in their efforts to understand what makes a 'good' and accurate recycler and to target resources for those who do not comply. On this issue, there was considerable dissent from the social scientists, who reiterated the point that the use of this kind of terminology was problematic and could result in further stigmatisation of marginalised individuals.

Data metrics

The GMWDA relies on specific kinds of data flows of the tonnage data concerning the materials which are collected. Every year, a total of 36,000 individual tickets are created and 900,000 tons

of waste are collected. They use this data to map habits, project trends and calculate recycling rates. Each month, the GMWDA use tonnage information to calculate the performance in each of the nine collection authorities. The information is inputted onto Defra's Waste Data Flow, a mandatory online national survey which is used to make the service delivery visible and operates as evidence that the contract with the private partner Viridor Laing (Greater Manchester) Limited, is being realised and maintained. The tonnage data is large in volume but does not classify as 'big data' as it does not fit the criteria in relation to velocity or variety. The authority is interested in whether they could use other forms of real-time data alongside this tonnage information to make better policy decisions, deliver more efficient public services and reach new EU targets.

New metrics

The now ubiquitous use of digital technologies in their diverse forms increasingly allows all kinds of activities to become informational. Data is routinely and continuously generated by devices, algorithms, instruments, and platforms that track transactions, or that elicit, classify and codify data in particular ways. Transactions or fluctuations can be tracked, counted, and modelled. The exponential increase in computing power and storage capacity has radically changed the availability of data, its modes of circulation, the means by which it is generated and presented and the ways in which it is used to model and predict future conditions in an ever growing number of fields. Rather than producing results compiled by recognised methods such as surveys, Big Data offers the basis for experimental projects using modelling and simulation techniques. In this way Big Data offers opportunities for analysing trends and patterns from aggregated data tracked in real time. It also offers opportunities for creating accurate forecasts, models and projections. Big Data sources therefore have the potential to change what is measured and how value is attributed.

Bolton has introduced in-cab technology to collect more detailed and timely information about their collection rounds. The introduction of this new technology signals a transformation from a paper to a high tech digital mode of working and therefore, a significant change in working cultures. The drivers' knowledge is crucial to deliver a service which caters for the needs of each local area. In Bolton, a tension has arisen between knowledge provided on the 'expert system' and the expert knowledge of the drivers. The Authority is interested in whether they could generate data continuously using other forms of digital technologies which could predict patterns and habits. For example, they would like to use sensors to record the quantity and quality of waste in bins and therefore predict future waste arising. The data could be used to create tailored services rather than relying on pre-designed routes. Rather than data being used for reactive monitoring this kind of Big Data would be anticipatory and predictive. It would open up the possibility of modelling, simulation and forecasting. This will be particularly important in the future.

In 2014, the EU introduced a new Waste and Circular Economy Package which is expected to change the way in which waste data is calculated so that a single methodology will be used across Europe to define the success of recycling rates and to gain clearer information on the quality of materials being recovered and circulated in the economy. The EU will require Member States to carry out precise measurement and reporting of what is actually recycled, as opposed to what is prepared for recycling. Their goal is to have 50% of municipal waste being recycled by 2020. The new form of measurement will require waste disposal authorities to record the quantity of materials recovered for recycling rather than the quantity of materials diverted from landfill which means that new forms of measurement will be required. Recycled materials are increasingly commodities which are being traded in the global commodity markets.

Re-use

Re-use is also an important element of the GMWDA's strategy. There are key moments where individuals throw away large quantities of waste, such as moving house or after the death of a relative. Items are usually passed along a social network or they are disposed of. The 'lumpiness' which is caused by these incidents is usually written out of large data models. Existing research questions whether the individual household is the most useful unit of analysis as materials flow through social networks. Could these unexpected moments of social life be accounted for in the data? Further, is the household the most useful unit of analysis? Would it be more appropriate to focus on an alternative measure to explore behaviour change such as the individual or the social network?

Big Data analytics

Traditional forms of statistical modelling and prediction depend on underlying regularity whereas Big Data involves new forms of uncertainty. The tools which are used are open-ended and serendipitous. It is not possible to know what you will find in advance and whether the data will be useful. The general question arises of how to ensure meaningful use and interpretation of the data. Rather than looking for averages, Big Data can be used to create patterns and predictions in order to understand populations in new ways. However, the experiments and simulations must be approached with caution as connections may be found in data but these are not necessarily causal or meaningful relations. There is so much data on offer that 'false correlations' are likely to emerge, and such 'findings' need to be verified by other means. Also, the science of forecasting depends on underlying regularities. In order to create accurate predictions data must be compatible. The question arises, would it be possible to predict current and projected data trends in waste management?

With Big Data there are outstanding issues of how to address the huge volumes of data generated, the advantages and disadvantages of techniques of mining and sampling, of

generating patterns, and identifying trends and correlations. Visualisations offer a powerful way of displaying data to look for comparisons, correlations and causation. They can be used as a space in which to articulate problems and explore solutions and may be useful when we do not necessarily know what the specific questions we want to explore. Also, urban laboratories could be useful experimental spaces for evidence-based research providing the basis for new collaborations between public bodies and academics that might help to negotiate the tensions between the need to 'make things happen' and the potential, but uncertain, benefits of exploring possibilities in an open-ended way.

Openness and ethics

Bins are not exclusive property. Rather they and their contents are public/private collaborations which connect the householder, the waste collection authority and other agents. If new ways of measuring data, such as sensors, are introduced the relationship between the public/private may be changed in interesting ways. Senior Officers of the Authority believe that there is a need for debating the potential of installing chips in bins as they believe that they would enable them to provide a more efficient and personalised service, but there have been a number of campaigns in the media which stress that chips or sensors in bins would be detrimental to the public, akin to surveillance technology which would allow the authority to 'spy' on households. However, it is questioned by social scientists whether the data which is generated from these devices would produce Big Data, or whether the data would simply correspond to existing metrics.

Feedback to the public

There is a high level of mistrust about putting sensors into bins and individuals' information being misused but there are other possibilities about what waste data could do. There have been a number of trials where a points system (similar to supermarket nectar points) has been introduced to incentivise recycling. With this technology, it is possible for individuals to 'opt in' and trace their personal information and thereby support an ethic of participation. Could chips or sensors on bins support the idea of ownership of bins, so individuals could trace their bin's contents? If data could be used to personal the service and make savings, could the sense of fear and paranoia about data sources be overcome?

Repurposing of data

Big Data challenges linear models of knowledge production and the ways in which information can be used to make decisions. Often data is re-purposed. It is connected with other data sets and therefore used for a different purpose than the one it was originally collected which raises questions about if data is made open, what could it connect to? Some members of the public are suspicious about information about the contents of their bin being collected and shared. Data protection is about limiting connections that the data makes and the area is fraught with

difficulties as it is often difficult to know how data will be repurposed. In Scandinavian countries new sensor devices have been introduced as there issues about trust are approached differently which signals a different relationship between the authority and the public.

Local authorities are under increased financial pressure and have less staff than ever before. Could private sector companies assist them and devise new apps with their data? Or, would it be possible for the GMWDA to learn from private sector companies and to use individual's data while ensuring that personal details are kept private? Even though there was clear interest and support in the run-up to the collaboratory from people working in the private waste sector, they did not attend at short notice. This meant that our discussions were limited in relation to these issues and the questions remains, how could we engage with the private sector on Big Data questions?