Urban Big Data Centre

An ESRC Data Investment

A Workshop Narrative Report: Interdisciplinary data resources to address the challenges of Urban Living

Organising committee: Jinhyun Hong (Jinhyun.Hong@glasgow.ac.uk); Mike Osborne (Michael.Osborne@glasgow.ac.uk); Iadh Ounis (Iadh.Ounis@glasgow.ac.uk); Craig Macdonald (Craig.Macdonald@glasgow.ac.uk); Joemon Jose (Joemon.Jose@glasgow.ac.uk); Mark Livingston (Mark.Livingston@glasgow.ac.uk); Catherine Lido (Catherine.Lido@glasgow.ac.uk); Katarzyna Sila-Nowicka (Katarzyna.Sila-Nowicka@glasgow.ac.uk)

Workshop Assistants: Sarah Currier (Sarah.Currier@glasgow.ac.uk); Alison Macgregor (Alison.Macgregor@glasgow.ac.uk); Keith Maynard (Keith.Maynard@glasgow.ac.uk)

EPSRC Institutional Sponsorship Funding awarded to: Prof. Vonu Thakuriah (Piyushimita.Thakuriah@glasgow.ac.uk) and Prof. Peter Triantafillou (Peter.Triantafillou@glasgow.ac.uk)

Dates: Monday 4th – Tuesday 5th April 2016

Venue: Sir Alwyn Williams Building (Level 5), University of Glasgow

Workshop Aims/Objectives:

1) To bring together an interdisciplinary mix of experience from academia, industry and the public sector to discuss perspectives surrounding emerging data (including Big Data).
2) To disseminate the data products and services available through the iMCD project.
3) To support network building and knowledge exchange between stakeholders.
4) To roadmap the future of big/complex data in the urban environment and disseminate these discussions in a position paper.

Report by: Keith Maynard

Introduction

On the 4th and 5th of April, 2016, The Urban Big Data Centre (UBDC), University of Glasgow hosted an EPSRC-funded 1.5 day workshop to explore the methodological challenges and innovations in urban data and to introduce UBDC's exciting new Integrated Multimedia City Data project (iMCD). The workshop, entitled '*Interdisciplinary Data Resources to Address the Challenges of Urban Living'*, was attended by a group of around fifty individuals comprising academics involved in collecting and using these novel sources of urban data, as well as methodological researchers, and public and private sector representatives of end-users. Also in attendance was a member of ESRC, demonstrating cross-council engagement in the area of urban living. The event was arranged into a number of distinct sessions in order to introduce and highlight the wide variety of applications and ongoing work relating to urban data and to stimulate discussion on subjects including end use, interdisciplinary collaboration and issues of privacy and ethics.

The morning of the first day served as an overview of urban data and the emerging role of Big Data in a variety of fields of research, with presentations from a range of invited speakers. The opportunity was then provided for all attendees to introduce themselves, voice their interest in urban data and what they hoped to achieve throughout the course of the workshop. A panel discussion followed in the afternoon, allowing a selection of people from private and public sector positions to consider how innovations in research into city data could benefit aspects of city planning, policy making and business improvement. The final activity of the day saw the attendees forming small groups in a breakout session to each discuss data applications and practices within the context of a certain theme. The second day commenced with short presentations from a member of each of the first day's breakout groups which fed into a subsequent discussion amongst a panel representing the Scottish Government and several world-leading academic institutions. The workshop concluded with a summary of its outcomes, the big questions that emerged and how they could be developed to form papers for publishing in an appropriate journal.

The workshop facilitated numerous networking opportunities during its course through collaborative exercises, coffee and lunch breaks and a group dinner on the evening of the first day. Furthermore, there was on-going commentary on Twitter throughout the event using the hashtag '#UBDCUrbanliving' to chronicle the days' proceedings and encourage engagement with interested outside parties, which can be viewed on the #UBDCUrbanLiving Storify.

The following sections of this report will attempt to summarise the content of the individual segments of the workshop.

Day 1:

Workshop Introduction

Professor Vonu Thakuriah (Director, UBDC) opened the workshop with a presentation to give some background context as to why there is increasing interest in Big Data and urban informatics. She argued that there is a desire for city managers to look for 21st Century innovation to help describe the changing social, physical and economic environment. Competing for investment and forming policy involves juggling a lot of elements within these areas and in short turnaround times, particularly at times of crisis. She noted that there are numerous challenges associated with current data solutions e.g. quality, level of detail, skill of users, and hardware/software availability and explained the value of leveraging Big Data as a vital step in measuring and improving Urban Living. She cited many examples of emerging sources of Big Data in the urban context, with particular emphasis on modern ways of collecting data using sensor systems and user-generated content, but many other sources such as customer transaction or administrative data, and explained that this wealth of new data offers a more data-intensive approach to help visualise, simulate and understand urban areas, and to take timely meaningful decisions. The presentations to follow later in the day would expand on these approaches.

This led her on to introducing UBDC's new ESRC-funded project, Integrated Multimedia City Data (iMCD), which is made up of a highly interdisciplinary team of Urban Studies, Computer Science, Education, Engineering, Geography and GiScience members. Promoting the many possible applications of this novel source of multimodal data that were collected and generated throughout this project was a particular incentive for arranging this workshop, with many attendees having previously expressed an interest or applied to use it. She announced that some strands of the iMCD is now ready to be used and that UBDC would be happy to discuss its potential with anyone interested in using it. One of the breakout groups later in the day would act as a further overview of the project and be led by Project Manager Mark Livingston. The data covers the greater Glasgow area with very diverse strands of data e.g. a detailed household survey (covering questions on demographics, employment, housing, transport, education, health etc.), activity tracking via GPS and 'lifelogging', extraction of news and social media data, satellite and other remote sensing data, traffic and weather. This was all collected over the same time period and to the greatest extent possible and so gives an exciting opportunity to view an operational city from many different angles. Vonu provided some examples of UDBC's ongoing work with the iMCD data; changing land use patterns and learning engagement in the city among older adults. Vonu ended her introduction by acknowledging the myriad of challenges associated in working with these large, unstructured datasets and the uncertainties and biases in the methods of collection before giving a general breakdown of the workshop ahead.

Presentation: *"Emerging forms of Data and Analytics"* – Prof. David De Roure, Director, University of Oxford E-Research Centre

Following the introduction, Professor David De Roure (Professor of e-Research at the University of Oxford) gave a presentation to provide a backdrop to new forms of data and real time analytics. He began by commenting on the increasing interdisciplinarity in his network of researchers turning to Big Data. He defined Big Data as being a 'deluge' of data so large that new methods need to be developed to cope with it. He added that rather than this new influx of data simply allowing us to

assess changes based on how we have done things in the past, it provides opportunities to look at entirely new research questions.

He began by describing emerging data in the form of social media generated data. He made particular reference to negative perceptions surrounding the governance of responsible use of real-time personal data, citing a recent MPs report[1] that labelled Twitter and Facebook terms and conditions as being 'more complex than Shakespeare'. He added that the unreliable nature of social media data intermediaries can lead to issues surrounding 'reproducability' with instances of researchers using the same sources but reaching contrasting results. He was keen to stress however that these discussions should not cloud the opportunities that social media can provide and that the data should be used in combination with the larger ecosystem of emerging data rather than being unhelpfully singled out. David went on to describe the other new forms of data within the categories of Internet data (of which social media is one strand), tracking data for monitoring the movement of people and objects, and satellite data imaging.

Prof. De Roure next described the emergence of risk and assumptions surrounding the concept of the Internet of Things where billions of everyday objects are now connected to the internet and generate data. He mentioned PETRAS, a new EPSRC funded hub for addressing privacy, ethics and threats to security and the trade-offs that often need to be considered to retain efficiency and utility. He spoke specifically about the increased need for automation and machine-learning to deal with the volumes of data being produced and the changing behaviour and role of human engagement with these systems. He quoted Berners-Lee's definition of 'social machines' as people doing "the creative work and the machine [doing] the administration" to retain a sense of human empowerment through these rapid developments.

He returned to the concept of an ecosystem perspective for the remainder of the presentation with the idea of humans being in a community alongside machines and the effects this has on the physical world. Particularly in view of the reliability of scholarly outputs, he expressed concerns over the design and purpose of robotics and social machines being introduced into the ecosystem and that there has been insufficient discussion and risk-assessment into their impacts and how it can be managed.

'Lightning' presentations

The workshop participants; a range of academics from the UK and abroad, and members of local government and private companies set out their expectations for the workshop in an around-the-room introduction. There was then a series of 10 minute 'lightning' presentations from seven guest speakers among the group to illustrate the range of opportunities these new forms of data are providing in research. The presentations and subjects covered within were as follows:

- Prof. Joemon Jose, Professor of Information Retrieval, University of Glasgow (Joemon.Jose@glasgow.ac.uk); *"Lifelogging – Issues & Opportunities"*

[1] The Science and Technology Committee, *'Responsible Use of Data'*, House of Commons; 2014. http://www.publications.parliament.uk/pa/cm201415/cmselect/cmsctech/245/245.pdf

Joemon discussed research from 'Lifelogging data', data collected using portable devices that track behavioural activity. He spoke about his data collection using an autographer, a wearable camera that takes a constant stream of photographs (100-200 per hour) in combination with a GPS tracker and the methods for extracting 'key moments' from the images. He suggested possible applications of health monitoring and new ways of observing the population and the urban environment before addressing privacy issues associated with capturing images of people without permission.

- Dr Charisma Choudhury, Deputy-Director, Choice Modelling Centre and Lecturer in Transport Engineering & Emerging Economies, University of Leeds (C.F.Choudhury@leeds.ac.uk); *"Behaviour Modelling Using Emerging Data Sources"*

Charisma spoke about data generated through people's short-, medium- and long-term decision making and how it can help predict and influence demand for products, infrastructure and services. She discussed the advantages and disadvantages of emerging sources of this data (social media, smartphone use, satnav etc.) compared to traditional survey methods in terms of collection costs, user participation biases and frequencies of collection. She presented the concept of 'Data Fusion' for more effective and reliable behaviour modelling by incorporating supplementary user feedback to validate GPS and other collections of tracked data.

- Luca Maria Aiello, Yahoo (alucca@yahoo-inc.com); *"Sensory Mapping"*

Yahoo's Sensory mapping projects explore the concept of quantifying the central elements of what people perceive affects the quality of life in city living. Mr. Aiello talked about an exercise of asking people to rank images compared side by side of parts of a city over which they consider to be more beautiful. He then spoke about creating 'happy maps' by using a computer to analyse these images at a much larger scale and to help identify which features in particular contribute to an overall positive or negative perception. Expanding from this, he showed similar sensory experiments for understanding people's attitudes to urban smells and noise to create multi-layered maps (GoodCityLife.org) and how these factors can influence, for example, people's route choices around a city in favour of the shortest path.

- Rod Walpole, Scientific Computing Officer, UBDC (Rod.Walpole@glasgow.ac.uk); *"Spatial Urban Indicators"*

Mr Walpole described UBDC's work on building Spatial Urban Data Systems (SUDS), a database that currently contains a large number of separate urban indicators for 14 major UK cities relating to: transport, housing, education, deprivation and the environment. He spoke about the efforts in making the spatial resolution of the data of sufficient detail to be appropriate for use in city planning and policy making (currently at Census Output Area level) and making it as open and accessible as possible. He showed some maps of Glasgow that have been developed by UBDC to help visualise parts of this data in order to demonstrate its numerous applications e.g. travel to work areas within particular travel times, relative densities of the housing rental market, and predicting vulnerabilities to flooding and other possible hazards. He closed by mentioning that finding good use cases is a primary focus for the project and directed the group to the project's Open Data Portal and Open Geoserver for data access.

- Achille Fonzone, Lecturer in Transport Modelling, Transport Research Institute, Edinburgh Napier University (A.Fonzone@napier.ac.uk); *"Data to Make Passengers and Public Transport Intelligent"*

This presentation showcased Dr Fonzone's work on real-time data collections for making transport systems more 'intelligent' for both transport users and operators. He presented a case study that he carried out in Edinburgh to show what the real-time travel planning information sources that passengers used were and how they affected their actual travel choices. He used another study of transport use in London to analyse travel activity and purpose using Automatic Fare Collection (AFC) sources and was able to detect 'multimodal commuter groups', passengers using multiple transport methods. He made the interesting observation that during periods where there have been underground line closures, commuters have been forced to use the 'Boris bikes' with a percentage of them continuing to do so after the lines returned to normal service, demonstrating the long-term hysteresis of travel behaviour resulting from interventions (in this case, underground line closures). This highlighted the degree of unpredictability and even some positive unintended consequences that can occur in transport events and hence the challenges of modelling such a complex system.

- Prof. Iadh Ounis, Professor of Information Retrieval, University of Glasgow (Iadh.Ounis@glasgow.ac.uk); *"iMCD Textual Data Services"*

Prof Ounis gave an overview of another branch of research conducted by the iMCD project focusing on textual data. His team specialises in developing tools for text processing that have been used in the retrieval of information pertinent to transport, weather and popularly discussed news topics on Twitter. The project has collected approximately 18 months of continuous Twitter data from users in Glasgow and Iadh demonstrated ways to mine this data so that it can be aligned with a variety of urban research questions and provide insights into the city of Glasgow that has not previously been possible. This can be done using this data in combination with the other iMCD data collected during the same time period and he gave suggestions of such further studies including: What are the public's sentiments when using various forms of public transport and how is weather a factor in this? How do major news events affect the moods of the local population (using the example of the Scottish independence referendum)?

- Katarzyna Sila-Nowicka, Research Associate in Urban Methods, Modelling and Simulations, UBDC (Katarzyna.Sila-Nowicka@glasgow.ac.uk); *"Sensing Human Activity: Twitter Foursquare and GPS Data"*

Dr Sila-Nowicka presented the part of the iMCD project involved with sensor data, focusing mainly on the GPS data collection as part of a survey. She briefly described the processing techniques involved for semantic enrichment of the raw GPS data (7 million individual data points) through stages of cleaning, segmentation, identification and classification of places and activities. Through combining this with Twitter data as well as Foursquare and inflows of 2011 Census data, she was able to identify functional regions of Glasgow such as their use for social, business or residential purposes. She ended her presentation with a short video visualising this data to show the separate movements of men and women throughout the city over an entire week.

The talks prompted a variety of questions from the group throughout the exciting, interactive session. Questions included:

- How does carrying or wearing monitoring devices affect the behaviour of the person collecting the data and people around them?
- How do you get access to sufficient amounts of mobile and other personal data if it is reliant on user consent?
- How do we gauge the level of 'attachment' and 'loyalty' to a physical space that isn't necessarily perceived as conventionally beautiful?
- How are urban indicators derived for areas without clear definitions such as green space – by free accessibility?
- What can be done to see how public transport operations can affect accessibility and participation, for example, in access to education institutions?
- What does having a temporal component to the data add in terms of possible uses?

The presentations provoked other ideas that were explored after lunch in the afternoon's panel discussion.

Panel discussion 1: *'Cities and Data'*

- Anne Connolly (Chair), Strategic Adviser to the Chief Executive, Chief Executive's Office, Glasgow City Council
- Ailie Clarkson, Statistician, ScotXed Unit, Scottish Government
- Peter Lindgren, Chief Operating Officer, TravelAI
- David McPhee, Head of Business and Digital Analysis, Scottish Government
- Alex Ramage, Head of Management Information Systems, Transport Scotland
- Steven Ramage, Ramage Consulting and Visiting Professor at IFC
- Bill Smith, Transport Planning Director, Ch2M

The panel of people from a range of private and public sector positions was chaired by Anne Connolly from Glasgow City Council who led discussions covering the following subject areas:

- Importance of data for decision-making;
- How government agencies and private companies are using data and examples of data programmes and strategies they are developing around it including novel, emerging trends in this direction;
- Expected end-users, services and innovations;
- Main barriers and challenges towards making effective use of data and in seeing opportunities posed by data;
- Panel members' viewpoint on the way forward for data-intensive decision-making

The panel opened with a discussion on the different ways in which urban data infrastructures are being developed. Because of these differences, a critically important need arises to have a common data language for interdisciplinary research and effective sharing of work to yield greater "interoperability". To illustrate this, an example was given of accident prevention in transport planning and the need for transparent data sources to help inform decisions and prioritise

investment for the most effective research. It was suggested that research questions with a specific focus on highlighting urban change and determining the value of such changes is what is of most value to planners and policymakers. Privacy issues associated with making linkages across datasets was brought into the discussion with regard to education data. The point was made that currently in working with sensitive and potentially disclosive data, there must be careful consideration among a panel for each specific use case. This process is unlikely to change and so for the foreseeable future, will remain a considerable time factor in being able to use certain data types in combination with the many emerging data sources.

Another main topic to emerge from the discussion was raised about the shortcomings in knowledge and interpretation of data among people in certain areas of industry and government. It was suggested that we should think about ways of increasing their understanding so that the power of new forms of data and tools for collection can be better recognised. These concerns were also said to contribute to the underutilisation of the increasing wealth of data that is becoming available. The challenge from the Government's perspective is to identify and harness the best uses of data for robust policy-making in a busy and pressurised environment. Support is needed from the wider data community to relieve some of the burden with respect to the limitations of time and skills possessed by departments holding the data. The panel agreed that a better collaborative approach is required so that other sectors are aware of and can benefit from the knowledge generated from specialised focuses. This was exemplified with reference to work with local authorities on transport issues and the realisation of how closely it is interlinked with the health and economic wellbeing of a city.

Collaboration formed a major part of the ensuing conversation and especially the barriers surrounding the sharing of data. These issues included: difficulties relating to searching and accessing other people's databases, the variable quality of metadata, issues over ownership and pricing of datasets that can hamper 'openness', and, particularly from the viewpoint of the Government representatives; the fact that their data and data practices are more heavily scrutinised by the public means they feel more risk-averse to how open they can be.

The discussion concluded with the exploration of ways forward, and a greater level of connectivity was a common desire among the panel. Academia was pointed to as being instrumental in finding the best data linkages for answering emerging urban questions and for deploying skilled data scientists to help mobilise data in sectors or particular offices where skill is lacking. Methods of improvement to skill and knowledge among the general population with regards to working with data were pondered. It was suggested that introducing modern data concepts at an early stage of education would be one way of addressing this gap. Another interesting idea was to make available a 'Big Data masterclass' course that anyone could attend to give them the confidence to engage more with Big Data, an action that has previously shown to be successful in improving understanding of open data concepts. Finally it was argued that the negative public perceptions surrounding privacy and data protection need to be overcome to allow more freedom to progress with dormant datasets. This could be done by making more of an effort to simply emphasise what the Data Protection Act actually covers and its purpose. Another way to help combat public suspicions would be to engage better with the public over data issues and to explore better ways of presenting information to people without advanced data analytical skills such as simple mapping and other helpful visualisations.

A variety of questions to the panel asked what practical measures could be undertaken to make use of Big Data in testing the impacts of change to urban landscapes in terms of the environment, educational opportunities and physical land use. The merits of testing 'temporary changes' on areas of a city and then using city data to measure the results was proposed. One example of this being done successfully was given whereby a car park in central New York was repurposed as space for local businesses and cafes to use and the positive effects that that had in the local vicinity. Similarly, using city data in models for simulations where such physical changes would be too expensive or otherwise impractical was also broached as having valuable applications in the future.

Summary of panel discussion 1:

- Using common data standards, terminology and metadata is critical for more effective sharing of work and interdisciplinary research
- For policy decision-making, it would be helpful to have expert advice on which datasets are of most value for a particular application or planning process
- Despite the increasing volumes of data being produced, linking datasets for novel research questions will continue to need access approval from a panel which can be a slow process
- Gaps in knowledge surrounding the best ways to record and use data was identified as a particular barrier to progress. Sharing knowledge and providing training across sectors would help improve data capacity
- The Government holds potentially very valuable urban data but it is currently underutilised for reasons such as lack of time and resources to make it more publicly available and aversion to the risk of negative reprisals through its use
- Increasing efforts in public engagement could help debunk some myths about data security and data protection and allow more freedom to use certain personal information for urban improvements
- Skilled data scientists could help advise on and 'mobilise' data in industry sectors where there is limited open data available

Breakout Groups

The final exercise of the day had participants form small groups to discuss data applications and practices. Each group focused on a particular theme with the outcomes to be reported the next day.

Day 2:

Breakout group presentations

A nominated member of each of the breakout groups gave a short presentation and described the main emergent points surrounding the following themes:

(1) *"Data sources, discovery and integration"*- Sarah Currier (Senior Project Manager, UBDC) spoke about underuse of data in terms of 'The data journey' and the challenges at each of the stages:

- Data sourcing – the approach is not always straightforward. With either designing research questions based on what data's available or establishing a research question first and then having to source the data, there can often be a circular process of refinements and compromises.
- Data acquisition – unwillingness to hand over data between organisations due to lack of time and skills, inadequate systems, anxieties about low data quality reflecting poorly on the organisation, privacy and ethical reservations, IP & rights
- Data storage and curation – more data being generated than can be stored and organised effectively both in physical storage space and skilled personnel available
- Data description – hard to get agreement on metadata standards on very heterogeneous data types and user groups. For Big Data, describing and categorisation is hugely resource-intensive
- Data discovery – because of the lack of consistency above, searching for applicable data is difficult as is making reliable data linkages

(2) *"Privacy, security and responsible innovation"* - Caitlin Cottrill (Lecturer, University of Aberdeen) raised the concerns brought about by a lack of a common working understanding of data privacy and security and considered ways to balance privacy with innovation. Some key points:

- De-identifying and anonymising are two very different things – there needs to be understanding that removing disclosive data fields does not make a dataset immune to allowing identification of individuals – vulnerability through data linkage must be recognised.
- Data responsibility – what is the balance of responsibility between private companies collecting personal information and consumers ensuring they're 'digitally literate' enough to be able to control what personal information they are agreeing to provide? Are there better ways of communicating privacy rules with the user?
- How can we keep up with effective data security encryption methods and govern the ethics of data and technology advancement in view of emerging robotics, the Internet of Things, wearable technologies?
- Are we forecasting change and training people sufficiently to be prepared for how systems will be in the next 20-30 years. How will the threats and data hacks shape our habits of sharing data in the future?

(3) *"Analytics methods and applications development"*- David McArthur (Lecturer in Transport Studies, UBDC)'s group discussion dealt with the challenges associated with becoming more

multi-disciplinary in approaches to methods for data collection and applications and made the following observations:

- Confusion could be caused by different disciplines using the same terms such as 'linear regression' in their viewing of data from very different angles.
- How are biases in methods of data collection being accounted for in determining causality? Geo-located Twitter data, for example, only represents 0.98% of tweets, what can be inferred about the population from such small proportions?
- Can accumulating much larger volumes of data make up for certain biases compared to a small number of records collected under stricter conditions from a well-designed survey? It depends on our aim – are we trying to prove a hypothesis or generate a basis for exploring new research questions?
- It is difficult to completely validate analytical methods without a 'gold standard' to compare to.

(4) *"Public engagement, citizen science and civic participation"*- Andrew McHugh (Senior IT and Data Services Manager, UBDC) led a smaller group than others as they tackled areas of urban data collection that involve direct public engagement. They divided the subject as follows:

- Principles of Engagement – How do we translate statistics and academic outputs to a mainstream audience? We need to be creative in methods of presenting information in relatable forms as a means of beginning a conversation.
- Citizen Science – Is amateur research and crowdsourcing credible and trustworthy? Can we use it to harness enthusiasm around activities and policy-making and shed light on issues within a community for further study?
- Civic Participation – A way to directly involve the public in decision making and feed the reasons for policy choices back to individuals they may affect in a transparent and dynamic way.
- End Users and Data Gathering – Can we employ a more explicit 'feedback loop' between data collector and provider to keep people central to the research without risk of data contamination?
- Trust – Public approval of research is valuable in its validation and improving its reach. To achieve this, trust needs to be developed through supplying information in a consistent, comprehensible and impartial way.

(5) *"iMCD introduction and FAQs"* - Mark Livingston (iMCD Project Manager, UBDC) – this group discussion was of a different format to the others and served to give a more detailed introduction to iMCD as well as providing an opportunity for participants interested in using the data to ask questions to members of the project. In addition, it allowed the group to focus on privacy issues using iMCD as a specific example to discuss the methodological approaches to anonymising data e.g. face-blurring of photos, linking GPS trace data to survey data, and exploring to what level of detail can the data be openly provided without individuals becoming identifiable.

Panel discussion 2: *'Analytical and Methodological Innovations for Emerging Forms of Data'*

- Prof. Marian Scott (Chair), Professor of Environmental Statistics, University of Glasgow
- Dr. Caitlin Cottrill, Lecturer in Geography & Environment, University of Aberdeen
- Prof. Ewan Klein, Professor in School of Informatics, University of Edinburgh
- Prof. Gwilym Pryce, Professor of Urban Economics & Social Statistics and Director of The Sheffield Methods Institute, University of Sheffield
- Gerard Scullion, Analytical Services Digital Manager, Scottish Government
- Prof. Stefan van der Spek, Director of education Geomatics, Delft University of Technology
- Prof. Robert Wright, Professor of Economics, University of Strathclyde

The workshop's second panel discussion was chaired by Prof. Marian Scott, UBDC Co-investigator, and featured representation from a range of academic institutions and the Digital Manager of the Scottish Government's statistical services division. It began by each member of the panel choosing to elaborate on one of the following topics:

- Changing nature of data collection and implications for methods;
- Quality, reliability, biases and uncertainty in such data;
- Challenges posed by the changing privacy landscape, responsible innovation and emergent ethics;
- Epistemological challenges to deriving knowledge;
- New lines of cities research likely to result from emerging sources of data;
- Utility and linkages to urban management and societal grand challenges that these new data will be used to address

Prof. Robert Wright opened by explaining that in his five years working on the ESRC's Methods and Infrastructure committee, he found that success in funding bids had a large dependency on how well they answered the 'so what?' question. Obtaining funding for new data collection increasingly requires being able to produce a clear set of underlying problems to address. He also made the point that when talking about Big Data, it is easy to lose focus on ensuring that the large volumes of data collected are still representative of the population we're interested in. He gave the example that he notices behavioural changes from other road users who extend him greater courtesy when he wears an obvious camera on his cycling helmet. For a dataset to be most valuable for secondary analysis, being able to prove what can be inferred and why it is better than other forms of data is what people such as policy-makers are looking for. Prof. Gwilym Pryce expanded on this sense of Big Data scepticism by focusing on the related societal challenges and particularly, the relationship between data efficiency and social inequality. He questioned the consequences of changes effected by the monitoring of very select groups through use of social media data and wondered how that could be mitigated. He said that he's observed three main arguments among social scientists to respond to, contributing to their current resistance to using Big Data in research: it is disconnected from theory, there is a fundamental mismatch between the purpose of data collections and what they are now being

used for, and uncertainty surrounding the understanding of results with respect to correlation and causation.

As commented by Prof. Ewan Klein, there is the perception that there is more data being collected than we know what to do with. In his experience however, there are many areas of urban research where data is either insufficient for his purposes or has not been collected at all. On a study into transportation in Edinburgh, he found there to be very little substantive data on cycling journeys and predicts that this type of data in the future will be relied on through voluntary reporting by individuals through smartphones and other such devices. Collections of this sort pose the questions of who do we trust to manage, analyse and further distribute this data. He floated the idea of having a mixed board of publicly elected 'data stewards' so that data can be trusted to be used for increased purposes by not having to immediately limit them through anonymisation. The concept of crowdsourcing limitations was also picked up on by Prof. Stefan van der Spek who agreed that GPS data alone, for example, while very detailed in some respects, doesn't provide much information about the user and type of activity. This means that correcting for error can be very difficult and so methods of collection for particular research purposes have to therefore be chosen very carefully. Likewise, he also commented on cities that are building their own 'data warehouses' and the temptation to compare cities using these large resources. He warned that because minor variations in methods of collection can have a big impact in what the data tells us, very similarly described datasets may in fact not be compatible.

Gerard Scullion, tasked with bringing together Scottish Government data in a consistent and unified online portal (statistics.gov.scot), spoke about the challenges of compiling and maintaining such divergent and continually expanding data platform. Reiterating the issues spoken about earlier in the workshop, he said that further work will need to be done towards creating helpful and searchable metadata to serve the multidisciplinary data community. Support from this community, he added, will be critical in movement towards making best use of the data available and advice will need to be taken to ensure that Big Data and its uses is adequately categorised and easily distinguishable from their other smaller datasets within the database.

Drawing from her teaching experience, Dr Caitlin Cottrill talked about a need for better ways for engaging students in order to equip them with a stronger skillset for using data. She had found that her students were being put off by the idea of coursework in statistics and economics but when she framed the subject matter using relatable examples such as in social media, they became more interested in learning about the value of this data and the questions raised by it. This led into a closing discussion between the panel, with contributions from the audience, about generating greater interest and enthusiasm for the potential of Big Data and avoiding its negative connotations. While it was recognised that its current prominence means that there are lots of funding opportunities available, it comes with the additional responsibility of having to prove its worth. Marian prompted the group for ideas of how we can avoid this narrow view and asked for suggestions of what we know to be the pressing urban data requirements. Responses included: addressing the significant absence of data collected in developing countries, demonstrating Big Data's utility in identifying and predicting environmental change, and exploring the increasingly detailed satellite images collected in the continuous monitoring of cities over the last 20 years and what new discoveries might be uncovered.

Summary of panel discussion 2:

- When talking about Big Data, it is easy to lose focus on ensuring that the large volumes of data collected are still representative of the population we're interested in
- One must be careful of what can be inferred from data when collected by methods that could influence response
- There is scepticism among social scientists surrounding Big Data for the above reasons as well as there being a sense that there is a fundamental mismatch between the original purpose of some data collection methods and what they are now being used for
- Crowdsourcing is an interesting new innovation for voluntarily collected data in areas never before possible. It does lead to questions of trust however in the reliability and comparability of the data as well as who can be trusted to manage the data.
- There is responsibility for people teaching data analysis and statistics in social science to do so in an engaging way to ensure future generations possess the skills for further advancement in the field of Big Data
- Categorising Big Data among other important data sources must be done clearly and consistently in new open data platforms
- More needs to be done to present the value of big data such as by addressing absences in data being collected in developing countries and demonstrating its utility in identifying and predicting environmental change

Workshop summary:

To close, Professor Vonu Thakuriah provided a summary of the workshop and drew attention to what she found to be the big questions that urban planners and decision-makers face:

- How to operate cities effectively and efficiently
- How should we evaluate potential consequences of complex social policy change on urban areas?
- How do we detect emerging trends in cities and regions?
- What makes the economy resilient and strong – how to develop shock-proof cities
- What interventions are needed for healthy and environmentally sustainable behaviour?
- What strategies are needed for lifelong learning, civic engagement and community participation?
- How does one generate hypothesis about historical evolution of social exclusion to impact current-day practice?

She then went on to summarise the key issues brought about by the various sessions. She highlighted the following points as being essential to our future research approach with regard to emerging urban data sources to make cities more effective and efficient:

- Timely detection of change in urban landscapes such that suitable interventions can occur on time to make a difference is fundamentally important to the future of urban big data.
- We must be careful not to oversell or dismiss available data sources. A multi-modal approach is central to understanding urban big data as the answer to a particular problem may not simply lie in one or two datasets.
- There has been an observed coming together of social scientists and computer scientists and it is crucial that this continues to develop in order to fill skill gaps. Working towards using a more standard nomenclature to describe data terms and methods is one measure for assisting this inter-disciplinary work.
- More attention needs to be paid to using Big Data for the purposes of modelling and simulation to predict and understand the future of cities.
- Citizen engagement and participation has a lot of future potential in data generation through crowdsourcing and volunteered geographic information as well as idea generation and urban decision involvement if interest can be effectively cultivated.
- Important that research and discussions around Big Data and other emerging data sources adequately consider or address issues such as biases, uncertainty, robustness of findings, and counter the surrounding scepticism.

Future Activity/Call for papers:

The workshop ended with an open discussion to set out the intentions for the publication of papers on themes covered in the workshop. There was a general consensus that, rather than attempt to cover everything, a small number of high-quality papers targeting a single journal would be the most realistic way forward. The agreed next steps were to try to converge on a particular high-ranking journal that will allow participants to target a clear title and organise their collaborations.

Appendix 1: Workshop Attendance List

Visiting Participants

Jonathan Brown, Glasgow City Council	Jonathan.Brown@glasgow.gov.uk
Andrew Burns, University of Glasgow	a.burns.3@research.gla.ac.uk
Chris Carlton, ESRC	christopher.carlton@esrc.ac.uk
Charisma Choudhury, University of Leeds	C.F.Choudhury@leeds.ac.uk
Ailie Clarkson, ScotXed Unit	Ailie.Clarkson@gov.scot
Anne Connolly, Glasgow City Council	Anne.Connolly@glasgow.gov.uk
Caitlin Cottrill, University of Aberdeen	c.cottrill@abdn.ac.uk
David DeRoure, University of Oxford	david.deroure@oerc.ox.ac.uk
Achille Fonzone, Edinburgh Napier University	A.Fonzone@napier.ac.uk
Lavinia Hirsu, University of Glasgow	l_hirsu@uncg.edu
Ewan Klein, University of Edinburgh	ewan@inf.ed.ac.uk
Peter Lindgren, Travel AI	peter@travelai.co.uk
Luca Maria Aiello, Yahoo Labs	alucca@yahoo-inc.com
David McPhee, Scottish Government	David.McPhee@gov.scot
Martin Miller, National Records Scotland	Martin.Miller@nrscotland.gov.uk
Laura Moss, University of Glasgow	laura.moss@glasgow.ac.uk
Gwilym Pryce, University of Sheffield	g.pryce@sheffield.ac.uk
Roushanak Rahmat, University of St Andrews	rr77@st-andrews.ac.uk
Alex Ramage, Transport Scotland	Alex.Ramage@transportscotland.gsi.gov.uk
Steven Ramage, Ramage Consulting	steven_ramage@outlook.com
Kate Reid, University of Glasgow	kate.reid@glasgow.ac.uk
Gerard Scullion, Scottish Government	Gerard.Scullion@gov.scot
Bill Smith, CH2M	Bill.Smith@ch2m.com
Stefan van der Spek, Delft University of Technology	S.C.vanderSpek@tudelft.nl
Bruce Whyte, Glasgow Centre for Public Health	Bruce.Whyte@glasgow.ac.uk
Robert Wright, University of Strathclyde	r.e.wright@strath.ac.uk

UBDC Participants

Orhan Aktas	o.aktas.1@research.gla.ac.uk
Graciela Carrillo	iil.gracielacarrillo@gmail.com
Chanialidis Charalampos	Charalampos.Chanialidis@glasgow.ac.uk
Sarah Currier	Sarah.Currier@glasgow.ac.uk
Jorge Gonzalez-Paule	j.gonzalez-paule.1@research.gla.ac.uk
Jinhyun Hong	Jinhyun.Hong@glasgow.ac.uk
Joemon Jose	Joemon.Jose@glasgow.ac.uk
Mark Livingston	mark.livingston@glasgow.ac.uk

Craig Macdonald	Craig.Macdonald@glasgow.ac.uk
David McArthur	David.Mcarthur@glasgow.ac.uk
Andrew McHugh	Andrew.McHugh@glasgow.ac.uk
Sean Moran	Sean.Moran@glasgow.ac.uk
Mike Osborne	Michael.Osborne@glasgow.ac.uk
Iadh Ounis	Iadh.Ounis@glasgow.ac.uk
Marian Scott	Marian.Scott@glasgow.ac.uk
Katarzyna Sila-Nowicka	Katarzyna.Sila-Nowicka@glasgow.ac.uk
Yeran Sun	yeran.sun@glasgow.ac.uk
Vonu Thakuriah	piyushimita.thakuriah@glasgow.ac.uk
Peter Triantafillou	Peter.Triantafillou@glasgow.ac.uk
Ashwini Venkatasubramaniam	a.venkatasubramaniam.1@research.gla.ac.uk
Rod Walpole	rod.walpole@glasgow.ac.uk
Yang Wang	yang.wang@hotmail.co.uk
Guoqiang (Chris) Wu	g.wu.1@research.gla.ac.uk
Jing Yao	jing.yao@glasgow.ac.uk

Workshop Assistants

Alison Macgregor	Alison.Macgregor@glasgow.ac.uk
Keith Maynard	keith.maynard@glasgow.ac.uk

Day 1 Breakout Groups

1	2	3	4	5
Sarah Currier*	Caitlin Cottril*	David McArthur*	Andrew McHugh*	Mark Livingston*
Orhan Aktas	David DeRoure	Jinhyun Hong	Anne Connolly	Mike Osborne
Sean Moran	Graciela Carrillo	Ashwini Venkatasubramaniam	Steven Ramage	Chris Carlton
Marian Scott	Peter Lindgren	Yang Wang		Katarzyna Sila-Nowicka
Jorge Gonzalez-Paule	David McPhee	Jing Yao		Yeran Sun
Ewan Klein	Ailie Clarkson	Chanialidis Charalampos		Achille Fonzone
Martin Miller		Iadh Ounis		Laura Moss
Roushanak Rahmat		Bruce Whyte		Kate Reid
Alex Ramage		Luca Maria Aiello		Guoqiang (Chris) Wu
		Stefan van der Spek		Charisma Choudhury

*denotes group leader