Big data and city living – what can it do for us?

Big data is transforming our cities. **Sallie Ann Keller**, **Steven E. Koonin** and **Stephanie Shipp** look at the benefits that big data can bring to society – and at some of the challenges as well.

Society is being changed by the huge amounts of data streaming from our cities every moment. Can human behaviour keep up?

What is big data and why is it important?

Think of all the major problems that currently face society. They range from obesity and health via national security and law enforcement to policy and commerce. They have one factor in common: human behaviour underlies them all.

Big data is bringing a revolution in science and technology. But it will bring a revolution still more profound when it interacts with people's decisions. That in turn will revolutionise the ordering of our cities, and of our lives. The social sciences such as economics, sociology, behavioural science and political science will be utterly transformed.

The world is facing a tsunami of data. It is coming from ever more ubiquitous sources, such as mobile technology, embedded sensors, blogs, social media, and location-based tools. See Box 1 for examples of where it is coming from and what it could do[1]. Big data is colliding with trends in evidence-based decision-making and understanding of human behaviour to create a revolution akin to the industrial revolution of the nineteenth century. How we prepare for this and embrace it will shape our future.

Big data will fundamentally change the way that social science is done. Physical sciences have been grappling with big data collected to study the physical world. In social sciences the paradigm will change, and much more radically because it involves the study of people and human behaviours. The traditional approach for social sciences begins with policy questions: you have a problem. It might be under-age drinking, or overcrowded transport or black-spots for crime. You commission a survey – probably small-scale, for reasons of time, practicality and cost – to provide some data about it; probably not a huge amount of data, but it will at least be relevant to your problem. Then you analyse the results and try to work out a policy in the light of them. In the era of big data this is too hit-or-miss and too inefficient to dominate. The new paradigm starts with data. It will collect all types of data – from all the sources in Box 1, and from many other sources as well – without knowing immediately how they might be put to use. It will save data that are collected for other purposes, such as cell phone or energy usage data, in hopes that it can be useful in answering new research questions as yet unthought of. No need now to go out and collect data when a problem arises: you already have enough – indeed, much more than enough.

It is true that the data is not now tailor-made for your question. It is not designed, as the traditionally sourced data are; we can call it organic[2]. Of course there

Box 1. The exponential growth of data: some examples and potential uses[1]

Organic data:

Location data

- Cell phone ”externals”
- EZ pass transponders
- Surveillance cameras

Political preferences

- Voter registration records
- Voting in primaries
- Political party contributions

Commercial information

- Credit card transactions
- Property sales
- Online searches
- Radio-frequency identification

Health information

- Electronic medical records
- Hospital admittances
- Devices to monitor vital signs
- Pharmacy sales

Other organic data

- Optical, infrared, and spectral imagery
- Meteorological (e.g., temperature, pressure, wind, humidity, visibility, composition)
- Seismic, acoustic
- Ionising radiation, biological and chemical

Designed data:

- Administrative data (e.g., tax records)
- Federal surveys
- Census of population
- Other data collected to answer specific policy questions

What might you do with this data?

Migration and location

- Measure urban migration
- Map population movements during natural disasters
- Identify neighbourhoods with inadequate social services
- Map human behaviour, such as dining-out habits, and correlate with health outcomes, such as diabetes

Transportation

- Optimise operations (e.g., traffic flow, utility loads)
- Develop rational infrastructure plans (e.g., zoning, public transit)
- Examine the distribution and patterns of health events (e.g., disease surveillance and screening)

Energy related

- Practice monitoring, reporting, and verification for greenhouse gas emissions treaties
- Detect hazards (e.g., leaks, plumes), emergency management
- Establish energy efficiency standards for buildings and appliances
- Use knowledge of behaviour to encourage energy efficiency

Methods and experiments

- Validate and calibrate proxies
- Conduct policy experiments and simulations
- Understand urban meteorology (e.g., leaks, plume dispersal)
- Synthesise large seismic apertures for seismology and earthquake engineering

will be no need to ignore the designer data if you can get it: designed data such as censuses and surveys will doubtless still be collected by the national statistical system and industry, and one trick under the new paradigm will be to combine designed data with organic. But by one estimate, organic data makes up 70% of all data generated and this amount is growing. For example, in 2011, there were 2.3 trillion text messages and 2.3 trillion minutes on wireless devices. By 2014, these counts are expected to increase by a factor of 35.[3] The amount of designed data is trivial in comparison.

Data storage requirements are predicted to grow to hundreds of petabytes by 2015; fortunately, costs of storing data are declining rapidly and new ways, such as cloud computing, are expanding. Less fortunately, our ability to analyse such quantities of data has not kept pace. Fully automated collection is a reality; fully automated analysis “is at best a distant reality”[4]. So although we may have data far in excess of our requirements, we shall only have analysis of the things that end-users actually need.

What will society gain from big data?

The big change to society will come from our ability to model how human beings behave. Social scientists are increasingly able to take a systems approach to understanding the impacts of technology on society. Patterns of behaviour, changes of behaviour – these will be more detectable under big data. The big challenges – energy security and availability, a sustainable environment, human health – involve complex interactions between the physical world and human behaviour. Dynamic and networked systems link the two. Big data will let us interpret and model those links.

Imagine knowing practically every detail about a city. The state of its infrastructure, its inhabitants, its environment are all known to you, to high resolutions in time and in space. You are able to fuse physical data streams with socio-economic ones: transport data tells you where people are going, sales and transaction data tells you what they are going to see or do or buy, tweets and social networks tell you how they feel about it – and tell you how those feelings change, minute by minute, as a few drops of rain fall or the sun comes out. (And of course high-quality weather data is already built into your system.) Suppose that practically every movement and action within the city's systems and infrastructure created another datum. Think of the data streams that would exist or could be created; the rates at which those data streams would flow; the technology and skills that would be necessary to acquire, store, and analyse such massive data. Think also of the theories and models that social scientists could generate and test, the problems that system operators and policy-makers could solve if they had access to those models and applications; and of the speed at which those problems could be addressed. Therein is the potential of big data. A city that worked like that could really get up and run.

New technologies offer opportunities to understand what goes on in cities in unprecedented detail and scope. The new data leads to a better understanding of the city's operation, and to new possibilities for optimisation of the infrastructure. Additionally, such data provides enormous opportunities to study general social phenomena, such as crime, entertainment patterns, or energy usage. The great techno-socio-economic beehive that is a city has become interpretable.

The concept of “smart” cities has sparked a great deal of interest around the world. Early successes in exploiting big data include policing in New York City and monitoring traffic in Singapore. However, the primary focus to date

© iStockphoto.com/adisa

has been on automation of transportation-related activities, such as the collection of data from toll roads. Another focus has been on energy tracking within buildings, usage of utilities, and integration of such data with financial data. As a city evolves and becomes smarter many other questions could be answered that would improve the quality of life and liveability within the city. These questions are largely unexplored by the social scientists studying cities, and for a simple reason: their traditions are not yet attuned to big data. The quality of organic data may be lower than traditional data provides, but corrections can be made given a proper statistical understanding of the data. Such approaches are already familiar in astronomy, remote sensing and biology. Their utility in the social sciences is just beginning to be explored.

So what social challenges can be solved with big data? The facile answer is "all of them". The problem is that, to solve these challenges, we must first develop better and more interdisciplinary ways of managing the data. Mixing structured or designed data (e.g., planning data, demographics and economic data) and unstructured or organic data (from maps, sensors and videos) is new to statisticians, and new ways must be found of combining them. Statisticians, like the social scientists, have a new world to get to grips with and a new, non-traditional set of skills to develop. Nor will their skills alone be enough. Teams working together, of statisticians, architects, urban planners, computer scientists, scientists and social scientists – all will be needed in the mix.

What are the challenges in using big data?

Along with its benefits, big data presents challenges. The amount of data is growing exponentially; the use of sensors and new technology is generating new types of data; and concerns are mounting about privacy and access. While broad access to data is a hallmark of democracies, accessibility could change as more private sector entities begin collecting and analysing data. Without the development of standards, agreements and infrastructures to share data, access to the data could be threatened, and the data itself could be unusable.

The financial and commercial sectors are already extracting value from the data. The issues in the social sciences are more complex because they involve human behaviour and people are slow to change – in the United States, it took 40 years to reduce by half the percentage of smokers and 30 years to increase seatbelt use significantly. (These timescales may be shortening: it has taken barely two decades for mobile phones to reach more than 85% of the world's population and in the process to become the world's most rapidly taken-up technology; and they have changed the behaviour of almost all their users.) Many of today's challenges can be related to how society approaches designing (or redesigning) our suburbs and cities, and our lifestyles, such as easing congestion, reducing obesity, and improving quality of life for an ageing population. In addition, economic conditions affect mobility and opportunity for change.

Some are sceptical of the ability to make increasingly specific forecasts about consumer behaviour. According to Peter Fader, co-director of the Wharton Customer Analytics Initiative at the University of Pennsylvania[5]:

> It reminds me a lot of what was going on 15 years ago with customer relationship management.… It turned out to be a great big information technology wild-goose chase. And I'm afraid we're heading down the same road with big data.… If you can get me a really granular view of data – for example, an individual's tweets and then that same individual's transactions, so I can see how they are interacting with each other – that's a whole other story. But that isn't what is happening. People are focusing on sexy social media stuff and pushing it much further than they should be.

A less sceptical outlook might be that the granular view – Fader's "whole other story" – is in fact the story that is happening. Time will tell – and possibly quite soon.

Institutional barriers, such as a lack of a data-driven mindset, can limit the effectiveness of big data as well. Coupled with this barrier is the shortage of individuals with the experience and education necessary to meet the expected demand for deep analytical positions.

To address privacy concerns, new approaches to making data available for research are needed that do not violate confidentiality of the entities in the data sets. In a related issue, the data must be available to researchers and not solely for use by companies or government. Balancing these issues while ensuring transparency will require new government regulations and incentives. While people may voluntarily give up their ownership of some of their data if they can see a benefit to themselves or to society, we are still in the process of determining the optimal level of privacy and consent. The smart city will be one where civic engagement is a reality: its people will participate in its running, its management, its ethos, and will feel part of that city. This may be the key to accessing data. As Gary King put it in a paper in *Science*: "Use and relevance are the most likely ways that the smart city will overcome resistance to openness, and skepticism."[1]

Retrieving and analysing increasingly massive amounts of data to inform decision-making will require new architecture and computational statistical theories and applications. Detecting patterns in data and creating outputs that can be used quickly and without ambiguity will become increasingly difficult as the data grows in size. The role of statisticians will become even more critical as traditional statistical methods are replaced by increasingly complex computational statistical approaches. Statisticians must provide the infrastructural and computational leadership that enables the use of big data in solving societal problems.

New York, Chicago and Seattle make data available to the public to spur creative ways of improving city life

Data from the Federal Statistical System is currently collected and used in 'silo-type' methods with one type of data and one type of data analytics addressing one problem. Yet the data likely to be most effective in providing a clear picture of society will include video, sensors, satellite imagery, electronic signals, and traditional demographic and social data that is unlikely to be fully exploited in any one silo. As few statisticians have the skills to handle this wide variety of data, what is needed is a "decathlete analyst" – human, machine, or human assisted by machine – able to process multiple feeds into a seamless unity. The days of an analyst for each type of data are passing[6].

What are next steps?

Many cities are seizing opportunities to make data available to all and to utilise such data to create smarter cities. For example, New York, Chicago and Seattle have made their increasing volumes of data available to the public. Their goal is to spur its use in creative ways that will optimise city life. This data can also be used to allow cities to generate insights about how events occur, to be proactive, and to improve their operations with respect to public safety, provision of health care, and optimising energy usage and traffic flow. These are only beginnings.

Long-term strategies are required. Just some of the needs are to develop sources of data and reporting criteria, and to motivate the collaborations and tasks that will be called upon to take on the big data challenge. It is critical that we begin to take these steps before significant precedents have been established that may restrict our ability to influence or implement optimal policies and procedures. We should act now to realise the full potential of the data-rich smart city.

References

1. King, G. (2011) Ensuring the data-rich future of the social sciences. *Science*, **331**(6018), 719–721.

2. Robert Groves, `http://blogs.census.gov/directorsblog/2011/05/designed-data-and-organic-data.html` (accessed July 12th, 2012).

3. Federal Communications Commission (2011) Annual report and analysis of competitive market conditions with respect to mobile wireless, including commercial mobile services. Available at `http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-11-103A1.pdf` (accessed June 12th, 2012).

4. JASON (2008) *Data Analysis Challenges*, JSR-08-142. McLean, VA: MITRE Corporation. Available at `http://www.fas.org/irp/agency/dod/jason/data.pdf` (accessed June 12th, 2012).

5. Gomes, L. (2012) Is there money in big data? *Technology Review*, May 3rd. Available at `http://www.technologyreview.com/business/40320/?nlid=nlbus&nld=2012-05-04` (accessed June 12th, 2012).

6. Sandra I. Erwin, Too much information, not enough intelligence, *National Defense Magazine*, May 2012.

Sallie Ann Keller is Vice-President Academic and Provost at University of Waterloo in Ontario, Canada, and the former Director of the IDA Science and Technology Policy Institute.

Steven E. Koonin, recently Under Secretary for Science at the US Department of Energy, is Director of the Center for Urban Science and Progress (CUSP) at New York University and an Adjunct Staff Member at the IDA Science and Technology Policy Institute.

Stephanie Shipp is a senior researcher at the IDA Science and Technology Policy Institute with an extensive background developing and analysing large federal surveys.