Chapter 8, forthcoming in Gerry Stoker and Mark Evans (eds*)*
Methods that Matter: Social Science and Evidence-Based Policymaking
(Bristol: The Policy Press, 2016).

‘Big data’ and policy learning

Patrick Dunleavy

In early February 2014, during an industrial dispute with management about extending the London Tube’s hours of service, many of the system’s train drivers went on strike. Millions of passengers had to make other arrangements. Many switched their journey patterns to avoid their normal lines and stations that were strike-hit, and to use those routes still running a service. Three economists downloaded all the data for the periods before and after the strike period from London’s pre-pay electronic travel card system (called the Oystercard), covering millions of journey patterns and linking each journey to a particular cardholder (Larcom et al, 2015). They found that one in twenty passengers changed their journey – an interesting ‘flexibility’ statistic on its own.

But after the strike they also found that a high proportion of these people also stayed with their new journey pattern when the service returned to normal, strongly suggesting that their new route was better for them than their old one had been. They considered two possible explanations of why people could have been using the ‘wrong’ Tube lines in the first place. One is that they were trying to maximize their welfare all along, but had limited their initial search behaviour because of high search costs, so failing to optimize. The other possibility is that Tube travellers only ‘satisfice’. They had not set out to maximize their welfare in the first place, but were just going with the first acceptable travel solution that they found. The scale of savings made by the strike-hit changers was so high, however, that only the second ‘satisficers’ explanation makes empirical sense. The analysts also showed that the travel-time gains made by the small share of commuters switching routes as a result of the Tube strike more than offset the economic costs to the vast majority (95%), who simply got disrupted. The unusual implication here is that economic welfare grew as a result of the strike. One implication might be that disruptions are always likely to have some side-benefits, which should be factored in by policy-makers when making future decisions (like whether to close a Tube line wholly in order to accomplish much-needed improvements).

This small case perfectly illustrates the huge advances in social science and public policy understanding that the availability of ‘big data’ now seems to offer. The economists could not possibly have reliably identified the sub-set of changing passengers from any conceivable survey of people in person. They needed a huge N of journeys, linked to specific Oyster cards (but anonymous as people), and completely objective and highly detailed in the information on routes chosen. The data involved was also not collected especially for this analysis. Instead it was routinely generated by London Transport as part of their administrative operations – checking that people were travelling with valid pre-pay cards with money on them, and at the end of journey debiting a fare electronically from each card. It is also noteworthy that all the data were digital from the outset, at the point where each Oyster card was read. For the analysis they were only recoded, not re-entered or handled manually at any stage. Finally, the process of understanding the data was swift. Within a relatively short time the analysis was complete and able to be presented in timely and accessible ways to policymakers (Larcom et al, 2015).

These are some of the features that mean the feasible scope of the modern social sciences could rapidly expand because of 'big data'. The societal scale of the effects may be large, as some 'pop social science' treatments have argued (Cukier and Mayer-Schonberger, 2013), and might even 'accelerate democracy' (McGinnis, 2013). But at this point in time the effects on public policy still seem to depend a great deal on the complex way in which a set of incentives and constraints on using 'big data' now play out.

The argument here is organised into four substantive sections and a conclusion. I first define what the term 'big data' means and consider where it fits within the already established 'tools of government'. Section two examines the varied and increasingly plentiful sources of big data, and considers how the phenomenon is linked with the digital revolution that is still working its way through many civil society institutions, especially government agencies and the public sector. I next consider how the social sciences' methods of analysis need to change as a result of big data's arrival. Section four looks at how 'big data' could alter public policy-making, and yet may not do so as much as one might think, because of barriers and time lags in its use. The brief conclusions situate these substantial but differently-facing implications within a far longer-run tendency of modern ICT changes to be simultaneously centralizing and decentralizing in their implications.

1. Defining 'big data'

The philosopher Rob Kitchin (2014a; and see Kitchin, 2014b) notes that what 'big data' means is still vigorously contested and debated (see Dumbill, 2012; Boyd, and Crawford. 2012), as with other new and fashionable tech vocabulary:

> 'Some definitions, whilst simple and clear – such as big data being any dataset that is too large to fit in an Excel spreadsheet – have limited and misleading utility [because] they do not get to the heart [of] what is different ontologically [i.e. in terms of existence] and epistemologically [i.e. in terms of knowledge] about big data. And there is a significance difference [here], which is why there is so much hype surrounding these data. For me, big data has [these] traits'… [in Table 1].

This approach helps explain why 'big data' is different from previous very large data sets, like the population censuses conducted every decade for more than a century in many advanced industrial states. Yet these huge exercises, of course, were the opposite of speedy or 'real time', often taking years to generate information. Like censuses 'big data' include whole populations, not any kind of sample – e.g. in the London Tube example, all journeys. But the

Table 1: Kitchin's definition of big data

▪ 'Big data' is 'huge in *volume*, consisting of terabytes or petabytes of data	▪ It is high in *velocity*, being created in or near real-time
▪ It is *exhaustive* in scope, striving to capture entire populations or systems, so that N = all	▪ 'Big data' is fine-grained in *resolution*, aiming to be as detailed as possible, and uniquely *indexical* in identification
▪ It is diverse in *variety* in type, being structured and unstructured in nature, and often temporally and spatially referenced	▪ 'Big data' is *relational* in nature, containing common fields that enable the conjoining of different datasets
▪ It is *flexible*, holding the traits of extensionality (so we can add new fields easily)	▪ It is *scalable* (so it can expand in size rapidly).

information is now collected and analysable in very timely ways, perhaps even in 'real time'. Because big data is very detailed and large-scale, and can often be linked to other information like GIS (geographical location) information, many more causal processes can be analysed in far more detail than is feasible with any survey method.

An alternative view of 'big data' argues that its essential features stem from being by-products of digital transactions. It is the digital properties of the data that make it easily re-combinable and scalable, which facilitates its retransmission, storage, and manipulability – all these are more general properties of many digital artefacts (Jensen et al, 2012; Anduiza et al, 2012; Constantiou and Kallinikos. 2014; and see Kallinikos et al, 2010; Kalinikos, 2006). According to Jensen et al (2012) digital artefacts, can be rendered in different archival structures without destroying the original artefact, which can be recovered later on. Unlike surveys or other forms of bureaucratic record-keeping, the 'big data' analysed are the native digital objects themselves (e.g. the journey 'traces' essential for operating the London pre-pay card system) rather than representations or codings of them. The data can be subject to loss-less copies and recombined without destroying the original.

The collection of digital objects is often relatively unobtrusive. For instance, commuters in the London Tube example knew only in a background way that their journey details were being recorded – and they had no chance of altering their 'data' as they could have done with a survey question. Numerous operationalizations and relations between data can be imposed without re-contacting subjects or otherwise producing response bias from iterated queries. Modern data storage capacities are also more scalable with big data, so that their collection and analysis can be automated, continuous and enjoined with other systems.

In public management terms, it is worth locating 'big data' (as defined above) against a somewhat modified and extended form of the 'toolkit' of government sketched by Hood and Margetts (2007**)** in Table 2. The key change I make here is to distinguish as two categories things that Hood and Margetts treated as one. These are:

- basic bureaucratic capabilities, of precise recording, classification of cases and information, reliable retrieval of information, and impartial and effective implementation – the *organization* infrastructure needed for any effective administrative apparatus in a basically Weberian mode; and

Table 2: The extended 'tools of government' (or NATOE) framework

Code	Label and resources included	Detectors (finding things out)	Effectors (getting things done)
N	**Nodality** – government's central location in civil society information networks	Citizen trust, generating civil society notifications to public agencies	Broadcast information and warnings, targeted messages
A	**Authority** – law, regulations, norms coercively enforced	Legal/regulatory requirements to report statistics or information	Prohibitions, tax raising and requisitions
T	**Treasure** – finance, property, conscripted resources	Tax-funded research and investigations	Subsidies, grants, tax exemptions, incentives, transfers, welfare state benefits
O	**Organization** – basic bureaucratic administrative competences	Maintaining an information collection network	Capacity to implement policies on the ground
E	**Expertise** – esoteric or highly developed technical knowledge and skills, organized in productive ways	Specialist scientific, research and analysis capabilities	Design, development and calibration of new scientific or engineering solutions or remedies

- the developed *expertise* needed if government is to do essential but inherently complex technical tasks – like keeping government IT safe from internet hackers, or determining safe limits for nuclear energy operations, or determining whether or not a new drug can be approved as safe. What is needed here is a combination of rarified professional talent, often extensive capital equipment, and highly sophisticated research/scientific types of organization.

The role of 'big data' within government is concentrated most in the expertise row (E), where later sections show that talent-acquisition and effective deployment are considerable problems. In highly professionalized systems, like health care, sophisticated data analysis can boost already established expertise in key ways (Bates et al, 2014; Moja et al. 2014; Sinsky et al, 2014.). It has also already opened up a potential for government agencies to develop genuinely 'free' (not just taxpayer-funded) services, where scalable information provision allows marginal consumers to be added at zero (or near-zero) marginal cost – a capacity that has transformed internet economics (Anderson, 2009). 'Big data' are also highly relevant in the nodality row (N here), where government must continuously compete in information terms with the major social and economic interests that it is seeking to regulate or influence. The state cannot afford to be blind-sided by better informed societal interests, for if this happens government's central role in society's information networks may be compromised or called into question. Monitoring participation and grievance-raising as it happens can also let government intervene proactively before problems get out of hand (Hale et al, 2014)

‘Big data’ can also contribute to the other three tools in diverse ways through

- improved basic organization (O); for instance, a border control agency upgrading its basic immigration scanning ability with biometric passports;
- better deployment of subsidies and grants (T); for instance, a welfare agency using data analytics to target transfers and employability help preventatively on people most at risk of becoming long-term welfare dependent; and
- improving regulatory capabilities (A); for instance, a police force adding the capacity to decrypt digital records kept by drug merchants on mobile phones or memory sticks; or gaining an ability to monitor social media being used by rioters to co-ordinate their behaviours, something London’s police failed to do in major August 2011 riots (Guardian, 2012) .

2. The main sources of ‘big data’

The massive amounts of new information becoming available to policy-makers come from two major sources:

A. ‘Administrative data’, a broad category including governmental records and tracking information but also data from commercial and business sources. And,
B. The ‘digital residues’ – the ‘electronic footprints’ of behaviour patterns, meanings and memes – created by our contemporary civilization.

Administrative data

A 2013 government Taskforce in the United Kingdom (UK) reported on the prospects for encouraging greater use by researchers of administrative data that:

> Government collects and holds a vast amount of data as part of the normal transaction of government business. Similarly, government collects data for the purpose of producing statistics about the current state of the economy and society. The ability to link and analyse data held by government has the potential to add new insights to our understanding of how society and the economy performs and to reduce the need for separate data collections where we ask, for statistical purposes, for the same information that has already been provided for administrative purposes.

Following a United Kingdom government commitment to improving data-sharing a rather complex apparatus for encouraging academic researchers to make greater use of administrative data was put in place, later on reorganized as the Administrative Data Research Network (ADRN, 2016). Most take-up of these new opportunities was at first by health researchers (perhaps 90%), followed later by education (perhaps 5%). More economics, or finance or transport-focused applications came much later on, in part because the ADRN framework is relatively restrictive and orientated a lot towards protecting individual identities.

Administrative data in government is all collected for transactional purposes, rather than being designed from the outset as a dataset for analysis by researchers (e.g. conducting repeat social science surveys), or forming part of the carefully constructed and evaluated national statistics reporting. Major longitudinal surveys are very expensive and large-scale social surveys, asking respondents for the same fixed grid of information over successive years, which has to be specified in advance. National statistics generally operate by requiring business or civil society organizations to file regular reports with government, filling in forms with numbers that are then checked and aggregated, e.g. to form a picture of the latest pace of economic activity in a country. Table 3 shows some pros and cons of administrative data set against these main alternatives. Administrative data typically records objective behaviours, but not people's opinions or meanings. It can be often collected unobtrusively and non-reactively (so you don't need to worry about people 'faking' survey responses or 'dressing up' statistics to try and make them look better. But the items recorded may not be exactly what research is interested in, but only indicators often or usually associated with what the focus of interest.

Some observers have identified 'missing data' as a problem of administrative data (Smith et al, 2004). But in fact technologically recorded data will often have less of a problem here than longitudinal surveys, since computerized recording devices can be very comprehensive in what they track (so long as they are working). However, tech-gathered data may often lack some central identifier, making it useful for some purposes but not others. For instance, police forces can be notified immediately of traffic jams by mobile phone companies when their data show long lines of geographically static phones forming along major highway locations, triggering real time warnings to other drivers and highway agency ameliorative action. But anonymized mobile phone numbers don't tell policy-makers where the car drivers (perhaps suffering repeated delays) have come from or are going to. Some more extensive data collection is often needed to get full value from administrative datasets.

One useful driver behind researchers' new interest in administrative data from government is that countries like the UK and Australia have made relatively firm commitments to *'open data' policies* – making available information already collected at taxpayers' expense so that it can be accessed and re-used by other actors in society, especially businesses, universities and civil society organizations (Margetts, 2013; Cabinet Office, 2013). The UK's site (at https://data.gov.uk/ accessed 15 February 2016) provides an overview of progress in this area and a wider openness agenda. Some work may be especially relevant for social scientists, e.g. on the National Information Infrastructure project (https://data.gov.uk/blog/progress-national-information-infrastructure-project accessed 15 February 2016).

The G7 group of countries have also formally pledged support for such policies, in the belief that the costs of data provision are already outweighed by direct benefits (Houghton, 2011) and that longer term it can help fuel innovation, especially by small and medium size enterprises. SMEs are often shut out of government sector IT contracts by huge integrated IT systems, giant contract sizes, demanding capability requirements and civil servants' precautionary desires to contract only with very large system integrator companies (Dunleavy et al, 2006). Yet smaller firms may have innovative and creative capabilities that are only weakly present in large public sector bodies or big system integrator companies.

Table 3: Some advantages and disadvantages of using administrative data in a 'big data' mode

Characteristic	**Advantages of administrative 'big data'**	**Disadvantages compared with other data sources (e.g. national statistics or longitudinal surveys)**
Scale	Large to massive scale collection. Often comprehensive for a whole population.	Analysts cannot over-sample those sub-populations of particular interest
Disaggregation by geographical area	Gives a reliably granular picture at the small area level	None
Frequency of updating	Updating occurs regularly (sometimes continuously) with all new transactions or contacts – usually annually, quarterly, or even more often	Updating is on an externally-fixed cycle and cannot be adjusted to capture specific events
Vulnerabilities	Achieving consistency in data reporting is a key compliance aspect for the managers and staff of agencies. Accuracy is required and inaccuracy may have seriously adverse consequences	Managers or staff may none the less 'massage' numbers where they can, so as to make their units' performance look better. Implicit knowledge combined with some space for discretion in classifications may make this hard to spot.
Quality checks	Managers check returns and data, focusing on case by case consistency. Internal audit will selectively highlight inconsistencies affecting performance.	Data quality is rarely cross-checked or tested using social science or statistical techniques or sophisticated data analytics – although external auditors may once in a while make more rigorous checks.
Metadata	May be limited or inconsistently applied across organizations	Later analysts may not have access to the implicit knowledge used in choosing metadata tags
Coverage of the population	Captures people who normally resist being included in conventional surveys Government incentives or coercion limit non-responses or incomplete data	Excludes people living 'off the grid' or not transacting with government agencies Coverage may vary over time if administrative rules change the costs and benefits for transactors
Data generated	Normally covers actual behaviours	Rarely captures intentions or perceptions The variables collected make sense for administrative reasons, but are not necessarily defined in useful ways for wider analysis.
Mode of collection	Less obtrusive than a separate survey. Penalties for misrepresentation and cross-checks of documents may enhance accurate factual data	Some reactive components (e.g. recall of historic factual data)
Identification	Machine-learning may let analysts compensate for missing registry links or identities	Beyond the original collection agency, most data may be available only in anonymized forms, not linked to key registers

Source: Own analysis but draws on Smith et al (2004).

A good example of the potential here again comes from London Transport, who struggled over many years to find ways of reliably communicating bus arrivals information to passengers waiting at bus stops, long after LT itself was able to track its buses' movements in geographical space (see Gammera et al, 2014, for an overall view of this problem). The official solutions focused on a centralized IT system that re-sent bus arrivals information to electronic signs at bus stops, not altogether successfully. The quite expensive shelters also had to be funded by advertisers installing the electronic link as part of displaying rolling or static adverts, a viable solution only for well-used bus-stops besides major roads. But once London Transport's real-time bus information was made available for outside programmers to access, it was not long before several competing small businesses developed mobile phone apps that told customers very accurately when buses would arrive at their own local bus stop – an especially useful facility for customers when it's raining or bad weather and where their bus stop has no shelter. Similarly the computerization of land registry data generated a huge potential for value-added services letting consumers know accurate house prices, organized successfully by a central government website in the UK (Land Registry, 2016) and (less well) by the private sector in the highly fragmented USA (Newcombe, 2014).

This example also illustrates a particularly strong and useful form of open data provision which occurs when government, business, academic or NGOs organizations commit to creating an *API or Application Programming Interface* where some of their 'big data' can be regularly (perhaps even continuously) downloaded in bulk for free and possibly in real time by other users (as with the bus geo-positioning information above). At other times the API grants other users access after a short delay that is key to safeguarding the data-collectors' commercial advantages, but may still be much better than official statistics' typically long time-lags. Providing APIs is especially valuable because researchers, businesses or NGOs have the assurance that the data stream will flow regularly and without interruption, and so they can reliably base their own operations on its accessibility. Of course, external re-users need a good deal of software and statistical expertise to download data from an API. But they do not need the time-consuming special permissions needed with much administrative data (e.g. accessing anonymized versions of people's tax or health records). In addition, APIs are very time-saving for expert users because they 'support code reuse, provide high-level abstractions that facilitate programming tasks, and help unify the programming experience' (Robillard, 2009, p. 27). They can also undertake re-analysis without having to let the primary data compiler know what they are doing (whether government or big private companies like Twitter or Facebook), and without having to alert the subjects of the analysis. Overall the shift to widespread use of open data and APIs 'allow[s] web communities to create an open architecture for sharing content and data between communities and applications. In this way, content that is created in one place can be dynamically posted and updated in multiple locations on the web' (Wikipedia, 2016a). A third way towards accessing 'big data' is feasible even when information is not being made available routinely in electronic form.

The technique of *web-scraping* allows researchers (typically expert journalists, academic researchers and PhD students) to visit public website displaying data or statistics in regular slots and a predictable fashion – as for instance many local authorities and health sector bodies do in reporting on their performance (Bradshaw, 2014)). A simple program is written that visits each of the relevant websites in turn and pulls off or 'scrapes' the same data in each, allowing it to be re-aggregated and analysed. Other aspects of public sector bodies' sites can also be recorded – e.g. how government ministries link to different internal and external organizations, recorded by Escher et al (2006) in a comparison of how foreign ministries in liberal democracies make use of digital communication capabilities.

Sometimes administrative data can only be accessed in anonymized ways, for privacy reasons or because of commercial restrictiveness (see below). So it often *lacks connections to key 'registers'* or key files that users within government do have access to. This 'incompleteness' is quite characteristic of administrative 'big data' and means that researchers have to be ingenious in finding new ways around the lack of name tags or unifying registers ID numbers. When plentiful data can be obtained that none the less lacks key registers, either social science theory connections or machine learning may offer a partial solution (see below).

Digital residues

It is a truism of our modern digital civilization that everything leaves a trace, an electronic footprint or record created as part and parcel of a massive flow of digital information (Siegler, 2010). Governments have not been slow to compel mobile phone companies and internet service providers to log and track every phone call and website visit or email in different ways, some focusing only on the metadata of linkages, others on their content as well. Despite some companies' resistance (Huffington Post, 2015), and efforts in Europe by the EU to guarantee a measure of citizen privacy (European Union, 2016), by and large the Snowden affair shows that all these records can then be hoovered up by intelligence and surveillance agencies such as the USA's National Security Agency, or Britain's GCHQ (Guardian, 2016; Snowden Surveillance Archive (2016). The agencies mainly use the data in hunting for evidence of terrorist activities and personnel or financial flows; or for information on criminal frauds, money laundering or illicit bank transfers. But with almost zero public accountability, concerns over indiscriminate accessing of metadata or message contents remain severe.

Away from such 'dark side' uses and concerns, however, digital residues are now often recoverable by ordinary social science researchers or by appropriately qualified teams working in public policy agencies. A first key source are *accessible text records* which have multiplied in the digital era. For example, the micro-blogging site Twitter has assumed huge centrality in the news and information systems of many liberal democracies, and the somewhat harder to access patterns of communication on Facebook are a huge repository of information previously mined for better-targeted advertising, but now available for far more multi-purpose analysis.

Greatly improved programs for text analysis mean that far more information can now be recaptured from text communications. For instance, stock market firms and financial reporting systems have long been analysing Twitter traffic relevant for financial markets using techniques of 'sentiment analysis' to try and detect turning points in the markets before they become obvious in changing behaviour patterns (Pang and Lee, 2008). The logic here is that if people become pessimistic economically they are more likely to sell shares and less likely to invest. So it is valuable for market actors to know that this is becoming an emergent feature of market trends *before* that pessimism converts into mass sell actions on the markets. Central banks have also begun to undertake somewhat similar text analyses, but this time on the lookout for emerging threats to financial stability or evasions of the regulatory net (Bholat et al, 2015). Similarly intelligence agencies increasingly find that increases in terrorist ‘chatter’ on aligned websites has some value in allowing predictions of major operations threatening security (Joachim, 2003). Finally, in the UK’s August 2011 urban riots, police forces like Manchester able to monitor would-be rioters’ chatter on Facebook and Blackberries (which are encrypted), and broadcast messages of their own on the same networks, did a lot better at preventative policing than those who lacked this capacity - like London’s Metropolitan Police (Guardian, 2012).

Mining masses of text for the relevant memes or distinctive ideas and vocabulary is now an essential aspect of many government’s efforts to maintain nodality. To stay in the centre of society’s information networks, government officials (and social science researchers assisting policy-makers) must increasingly be expert in analysing societal ‘big data’.

Nor are digital residues confined to text. Increasingly sophisticated systems now allow the massive recording of audio files from phones or other sources, and of images or videos from CCTV systems or other systems - such as the UK’s Automatic Number Plate Recognition (ANPR) system that tracks cars across main highways (Police UK, 2016). With storage very cheap (even almost ‘free now) search programmes can also scan for objects of interest either in real time - as with US cities’ programs that use facial recognition software to try and track the movements of known criminals via CCTV cameras (Gates, 2011, pp. 63-96); or in relatively swift retrospective (e.g. in criminal investigations).

Many of these developments mean that government agencies or social science researchers can often collect text or other digital data series that rather resemble administrative data in some aspects. For instance they may often be anonymized or lacking an authoritative register. But on the other hand they contain a great deal of potentially useful text, image or sound information, if only they can be decoded. We are still at the beginning of working out how these capabilities will affect the traditional Weberian model of government bureaucracy. Currently these agencies show an overwhelming emphasis on reducing all data about behaviours to highly parsimonious classification text (or now dots and dashes on computer discs). This stance reflects an inherited dominant assumption that storage is expensive and analysis will be relatively rarely undertaken. Yet bureaucratic form-filling loses tremendous amounts of information in the process of forcing people to condense answers in minute scraps of text. Why not instead use a videoed or audio-recorded interview with an agency script prompting for answers, from which key responses can be extracted via analysis programs – leaving the whole digital record intact for possible later re-analysis? Already Australia’s tax office (the ATO) employs digital voice recognition to grant or deny individuals access to their tax records, and employs voice analysis software in its contact centres to detect when people phoning their staff are under high stress (e.g. perhaps because they are lying about aspects of their tax affairs?). In future advanced bureaucracies may well record and store digitized video and audio transactions in loss-less ways.

3. Social science methods and the analysis of ‘big data’

The advent of ‘big data’ has created a great deal of uncertainty about the evolution of the social sciences from two major sources: 1) the growth of new methods and approaches, radically different in character from past approaches and 2) controversy about the continued importance of theory.

New methods and approaches

Many of the characteristics of 'big data' reviewed above mean that how it is analysed has to change radically. Very large N datasets can be processed far more effectively and extensively than conventional small N (under 5000 cases) survey datasets. In traditional social science even longitudinal surveys typically have too few cases to allow the intense dissection and multiple variables and categories that analysts often want to evaluate. And in regression or multi-variate analyses it is common to 'run out' of the variance needed to test complex models. In 'big data', by contrast, it is possible to focus analysis precisely, and to understand complex patterns of behaviour and factors affecting multiple, relatively small sub-groups of the population (like the 5% of Tube passengers who found better routes to work as a result of the exogenous shock of a strike).

Traditional social science with small N datasets also traditionally placed a lot of emphasis upon significance-testing as a way of knocking variables out of contention in the model building and model-testing sequences (American Statistical Association, 2016). Through convention, analysts looked for variables significant at a 5% probability (p) level in multi-variate model building, and regarded variables significant at a 1% p value as gold quality components (Wasserstein and Lazar, 2016; Radziwill, 2015). This approach has always been controversial, for five main reasons, neatly summed up and linked to five modern critics by Cook (2008):

> '1. Andrew Gelman: In reality, null hypotheses are nearly always false. Is drug A identically effective as drug B? Certainly not. You know before doing an experiment that there must be some difference that would show up given enough data.
>
> 2. Jim Berger: A small p-value means the data were unlikely under the null hypothesis. Maybe the data were just as unlikely under the alternative hypothesis. Comparisons of hypotheses should be conditional on the data.
>
> 3. Stephen Ziliak and Deirdra McCloskey: Statistical significance is not the same as scientific significance. The most important question for science is the size of an effect, not whether the effect exists.
>
> 4. William Gosset: Statistical error is only one component of real error, maybe a small component. When you actually conduct multiple experiments rather than speculate about hypothetical experiments, the variability of your data goes up.
>
> 5. John Ioannidis: Small p-values do not mean small probability of being wrong. In one review, 74% of studies with p-value 0.05 were found to be wrong. (And see Ionnides, 2005)'.

In big datasets the problems of significance testing simply disappear, however, because dozens of possible associations between variables (maybe even all or most associations) will pass the 1% or 5% standards, simply because the dataset Ns are so large. So significance testing is no use at all in winnowing or evaluating models – as perhaps it never should have been used. But how then do you distinguish a preferred model when so many variables can no longer be discarded?

In a paper on 'new tricks for econometrics', the Google chief economist Hal Varian (2014) outlined a different, competitive approach to model testing. A large dataset might be divided randomly into (say) ten parts and different model streams evaluated in different parts by their ability to predict the behaviours or patterning being researched. So model development here is competitive, and model performances are compared across different subsets of the same massive dataset. Empirical competition winnows out dud models and promotes high performing ones, not significance tests or theoretical considerations.

At root here, perhaps, is an approach that values 'control' over 'causation'. The traditional approach (often over-focusing on regressions) tries to develop a causally developed explanation of phenomena that are the focus of research, where the set of 'independent' variables are conceived as (somehow) *producing* the effects analysed, such that a rerun of the same factors in the same situations will reproduce the results being analysed Yet real predictions are rarely attempted (Taagepera, 2008). And, anyway, social situations never rerun from the same point (Gleick, 1987). As the Greek sophist Thrasymachus insisted (in a parody of an earlier dictum of Herodatus): 'You can never step in the same river once'. The alternative approach focuses on *control*, being able to manipulate 'explanatory' variables as effectively as feasible so as to produce desired results and interpretations. Given the complexity of social life, some analysts at least are now sceptical that causality can be uniquely established.

The control perspective suggests that the acid test of 'big data' analyses is not the backwards-looking interpretation of what has happened already – traditional social science's almost exclusive focus. Instead the digital and 'big data' context of much policy making and administrative implementation opens up scope for a particularly ambitious form of the experimental approach discussed in small N forms by Peter John in Chapter 4 below – randomized control trials or RCTs. *Online* RCTs with 'big data' have all the virtues that Peter discusses, but a lot fewer barriers and difficulties because the availability of huge datasets allows the evaluation of the very small-scale effects that are all we can realistically expect as a result of experimental stimuli. Online RCTs can also often be undertaken at low cost and in real time by government agencies or businesses.

For example, in the UK getting 1.9 million people a year to pay court fines promptly is an important aspect of the conduct of court services. Where fines are not paid promptly far greater expense is caused for the government, in chasing unpaid debts using contractors, and far greater costs also result for the offenders involved. Now people's willingness to pay may actually be influenced by quite small factors – such as the design of reminder letters sent to them, the briefness and clarity with which the consequences of non-paying are explained, and the ease or convenience of paying immediately (e.g. via credit card over the phone or online). An online RCT might come up with one or more treatments (such as one or more new and 'improved' or redesigned forms of reminder letter). These are sent to large, randomly assigned treatment groups, and their performance in improving prompt payment of fines is then compared with a randomly assigned control group. At the end of such an experiment we may know that treatment B works better than A or C, and generates a worthwhile saving for government finances compared with the status quo reminder letter sent to the control group. But we may still be none the wiser (beyond some intelligent guesswork) about exactly why letter format B works so well and other ideas less well.

An ambitious agenda claiming to resolve this problem is associated with 'behavioural public policy' studies and behavioural economics (much of which has actually be developed by psychologists. Their claim is that social science can now exhaustively explain the origins of dozens of different 'anomalies' or 'fallacies' that affect individual citizen or customer behaviours, influencing them to deviate from what abstract rational choice models predict that they should do. In the UK, the 'Behavioural Insights' team in partnership with the Cabinet Office claimed to use online or other 'big data' RCTs as part and parcel of generating comprehensive behavioural insights into why citizens sometime behave sub-optimally and how to 'nudge' people unobtrusively into making better choices through structuring how options are put to them (Halpern, 2015). This extended the well known Sunstein and Thaler (2009) arguments about how to 'nudge' people unobtrusively into making better choices through structuring how options are put to them (originally called 'paternalistic libertarianism'). Some social scientists have enthusiastically endorsed the potential for such techniques allied to experimentation to produce useful additions to our knowledge of 'what works' (for instance, John et al, 2013).

Yet the jury is still very much out on whether behavioural public policy is anything more than a developed set of descriptive narratives that can be cited in post hoc justification by analysts discovering that particular control effects work – exactly the way that 'behavioural insights' are used in marketing by private sector firms. Around 60 economic 'fallacies' have now been identified in behavioural economics and psychology, and there is an even larger list of 'cognitive biases' when social and memory biases are considered (Wikipedia, 2016b). So it is almost certain that several apparently relevant behavioural or other cognitive biases can be cited to 'explain' almost all effects discovered in big data and online RCTs, thus suggesting an ability to 'control' or shape behaviour in some way.

The track record of behavioural insights being applied in government and public policy contexts is also a relatively short one and it is not clear from most studies if 'Hawthorne effects' have been controlled – that is, 'treatment' effects arising from customers just meeting for the first time an innovative, different or experimental approach to policy development. So the first time we send people due to pay court fines an improved reminder letter, we may get a noticeable gain in the desired behaviour - more people pay promptly and without having to be expensively chased. But next year that 'improved' letter is the new normal. So will what worked last year shows results again, on repeat? The most likely scenario of 'behavioural insights' using 'big data' is that researchers or policy analysts within government get stuck on an escalator of introducing new innovations every period to counteract the 'wearing off' of past innovations, that become over-familiar. In other words, government will have to continuously 'market' public services to service users or regulatory targets, in almost exactly the same way as private firms must continuously use marketing techniques to attract customers' attention. In this scenario, we may get better at controlling or prompting citizens' behaviour, but not necessarily in understanding why. If so then 'behaviour science' peters out in the same kind and level of insights as private sector marketing, where the premium is on creatively stimulating clients with new or unfamiliar materials, not on building up a cumulative or well-founded body of knowledge.

Different kinds of problems may arise with the application of 'machine learning' techniques in big data contexts (Armstrong, 2015). This is a form of automated engineering software that copes with the 'incompleteness' of much big data by computers working through automatic algorithms that allow them to 'learn' from the associations of what variables they do have about other variables that they do not have. Over several or many iterations the software improves how it categorizes the information being handled, and works out better how to construct associated variables that can speak to the underlying identify of different groups of people, behaviours or assets covered in the data.

> 'Large datasets may allow for more flexible relationships than simple linear models. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, and so on may allow for more effective ways to model complex relationships' (Varian, 2014, p. 3).s.

For example, from a health 'big data' set we may not have access to people's individual identities. But by associating symptoms and disease problems together, the robot analysis may 'learn' a great deal about the characteristics of different groups and thus bridge across the non-availability of the identities. (This capability is one reason why releasing even anonymized health data set is still highly controversial).

Machine learning is 'concerned primarily with prediction', but it is also closely linked with data-mining, which primarily focuses on summarizing data and extracting interesting findings (Varian, 2014, p. 5). However, machine learning also has relevance for other areas, such as making estimations and improving hypothesis testing.

The continued importance of theory

In an interesting critique the US political scientist Nicholas Christakis (2013) pointed out that the structure of disciplines in the contemporary social sciences has endured remarkably unchanged for decades. By contrast, discipline-based and university departmental structures in the STEM sciences have changed radically (sometimes several times, as in biosciences) to reflect new methods and foci of study. Christakis argues that the development of new fields will (or at least should) produce similar changes. Key contemporary developments might be a shift to studying the influence of genetics on social behaviours, the advent of neuro-economics, or the development of 'big data' analysis towards software engineering and mathematical analysis (instead of past social science methods and packages adapted to far smaller N, survey datasets).

Elsewhere in the field of empirical sociology, some observers have speculated on the coming crisis produced by sociologists being locked out of many 'big data' sources by commercial confidentiality on the one hand in business examples, and by government privacy restrictions and administrative non-adaptability on the other (Bélanger and Crossler, 2011). The main casualty of the crisis is again likely to be survey-based work, which in many respects can no longer 'compete' with the insights that 'big data' owners can command:

> To give a simple example of the merits of routine transactional data over survey data, Amazon.com does not need to market its books by predicting, on the basis of inference from sample surveys, the social position of someone who buys any given book and then offering them other books to buy which they know on the basis of inference similar people also tend to buy. They have a much more powerful tool. They know exactly what other books are bought by people making any particular purchase, and hence they can immediately offer such books directly to other consumers when they make the same purchase (Savage and Burrows, 2007, p. 891).

Yet as the data universe changes rapidly, the authors also think that there are important theory and professional 'ideology' barriers to sociologists accessing many 'big data' sources. They see a 'danger [of sociology] taking refuge in the reassurance of our own internal world, our own assumed abilities to be more "sophisticated", and thereby I chose to ignore the huge swathes of "social data" that now proliferate' (Savage and Burrows, 2007, p. 887).

The wider debate here leads to the ambitious claims of the IT writer Chris Anderson (2008) that 'big data' ushers in 'the death of theory', because in the digital era everything in contention between rival schools of thought in social science can in principle be explored and tested, using our vastly expanded armoury of information and methods of analysing it:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves'.

Widely rejected by social scientists as almost half-baked in conception (Williamson, 2014), this simple prediction may none the less have a degree of force, certainly in the way that public policymakers approach research issues. A good example is the kind of 'predictive policing' programs developed by some US social scientists for cities like Los Angeles (Perry et al, 2013). They use machine learning and data mining techniques to hunt for data associations that will allow them to identify possible crime suspects or traffic accident problems, and then analyse the behaviours of people involved. From there the analyst can suggest to police patrols very specific locations and times that they should be in particular small 'boxes' of the city where known offenders were previously spotted and might be expected to return, or particular zones and times where the presence of police patrols may help deter drunk or drug driving. The theory behind these models is often pretty slender (e.g. offenders are creatures of habit and hence tend to revisit the same locales at the same times of day). But if they seem to work in offering better arrest records, or crime prevention/ deterrence effects, or impacts on cutting accidents, then policy administrators and politicians will still want to deploy them.

The mainstream social science rebuttal of such examples has been to argue that 'big data' analyses, run directly by software engineers or 'big data' owners, only explore or test 'common sense' kinds of propositions – in many cases validating the blindingly obvious, and failing to control for the inherent variability of social behaviours. A now classic example is the Google algorithm that for a period of some months and years successfully predicted where flu outbreaks would take place in America, using a big dataset of people's daily flu-related queries to the search engine (Ginsberg et al, 2009). For a time this even seemed to work faster and better than the US government's elaborate system for recording and notifying the incidence of diseases, run through the public health regulatory system (by the Centre for Disease Control). But the apparent association did not last – the advent of a different strain of flu made the Google-based model perform very poorly, forecasting almost twice the level of flu in 2012 than the CDC system found (Lazer et al, 2014). The official recording system proved resilient, while the 'big data' alternative had key problems. So the conventional wisdom now is to 'pooh-pooh' the 'death of theory' and to claim that the real future lies with 'big data' analyses informed both by strong theory as well as improved data analysis (Williamson, 2014). Whether such a view is well-founded remains to be seen.

4. Using 'big data' in policy-making

The examples already mentioned give a good sense of how 'big data' has already begun to be used in a wide range of public policy settings, both by social scientists and by professional analysts working for state agencies or consultants. But these 'classic' cases of impacts are also potentially unrepresentative of the public policy landscape as a whole. To give a broader picture it is worth briefly considering an admittedly anecdotal summary of central government departments and major agencies in the UK. The most advanced organizations in terms of using 'big data' seems to be the Government Communication Headquarters (GCHQ, the UK's electronic surveillance agency), which is also known to be using machine learning to assist in 'big data' searches. For other intelligence agencies, their scientific or analytic capabilities are less evident. Medical researchers in academia and the National Health Service (NHS) funded by the UK's Medical Research Council have also accounted for around 90 per cent of the applications to use administrative data, and for the large bulk of publications including mention of administrative data. The Government Data Service has also used machine learning to predict traffic flows on its major site and identify anomalous times or event periods (GDS, 2014).

Some large civil government agencies with lots of transactional data are using 'big data' analysis a fair amount to analyse policy problems, and have begun making their data accessible to researchers and using machine learning also a little bit. The UK's tax agency (Her Majesty's Revenue and Customs) and Transport for London are prominent here. But other departments with massive transactional data have only just begun to think through the issues and there data is still closed to most research – notably with the Department of Work and Pensions (the UK's social security and labour markets ministry), and the Ministry of Justice, which sits at the apex of the legal structures and runs prisons. Some 'big data' is being analysed around crime issues by the Home Office (responsible for police forces). But many Whitehall departments running substantial policy fields (like education, business and regulation, policing, transport and environment) have neither ready access to 'big data' resources of their own (except longitudinal surveys in some cases), nor do they have the highly skilled staff or developed and stable research contacts in universities to have yet mastered 'big data' analytic capabilities. Even highly numerate and analytically orientated departments operating in environments with plentiful 'big data', such as the Bank of England setting UK interest rates and running financial resilience regulation, have only just begun to explore 'big data' possibilities. They remain very dependent upon the national statistics system for the coverage and timing of their policy information (Bholat, 2014).

For much of government a key driver behind developing the capacity to do 'big data' analysis lies in the incessant competition between state agencies and private sector business, (and even civil society NGOs) over shaping public policy and regulatory interventions. If government is to retain its nodality, a central location in society's information networks, and if ministers and officials are to speak with authority on issues, then state agencies cannot afford to fall behind in their capacities to acquire and analyse timely information. For instance, the competition between regulators and regulatees is incessantly changing as regulated firms and industries constantly innovate in developing new products and services. In a recent study of *The Impact of the Social Sciences* Bastow et al (2014, p. 133) quote executives from a major IT company discussing how research can shape policy development:

> I don't know if we are getting ahead of universities. But we are getting ahead of the government, that's for sure. I was at a Treasury thing yesterday with another colleague, and we were talking about datasets and so on, and this guy from the Treasury was saying "That's all very well, but we survey 1000 people every week, and I feel pretty confident with that. How robust is your data?" And we were just like, "Well this graph here is based on 207,000 people from yesterday". So we are getting ahead there.

The 'arms race' character of government keeping up is also obvious in fields like police forces trying to keep pace with criminals and anti-social forces; and by the 'dark side' world of government versus terrorists, or security systems versus hackers. But it also applies far more widely innovations - for instance where regulatory agencies are supposed to be monitoring or controlling the constantly innovating activities of private sector firms in key sectors of the economy like financial institutions or stock market trading using computer algorithms (Government Office for Science (2012) .

Yet the barriers to large-scale deployment of 'big data' methods in government and in policy-related social sciences are also substantial. Despite the growth of APIs and government pledges to open up data to outside users, substantial privacy barriers remain in areas like medical and health research (where patient confidentiality is sacrosanct). In taxation also, citizen privacy is a key concern in the Anglo-American democracies, but in Norway personal tax submissions are public documents (see Devos and Zackrisson (2015) for these cases and a useful survey of global practice). Legal and constitutional protections for citizens and businesses have generally failed to keep pace with the capabilities for data surveillance by government intelligence agencies, which can broadly do what they want in many countries.

Yet still this whole area of citizen privacy remains littered with restrictive rules and regulators. Although normally insufficient in the UK or USA to provide citizens with worthwhile safeguards against state intrusion on their private lives, these barriers and provisions are none the less often enough to create major difficulties in releasing even anonymized data in sensitive areas. In addition, many different government agencies and large and small businesses have faced massive problems over either mass data records going missing or being lost, or in hackers proving able to access and download sensitive information on a large scale (Huffington Post (Australia), 2016). These have all added to the difficulties of moving to 'open data' arrangements and of getting researchers (or even other government agencies) access to many kinds of 'big data'.

The other substantial problem for civil government, and for university social science quite acutely, is that they tend to be at the back of the talent queue to recruit personnel with the right kinds of training and skills to be good at 'big data'. The social sciences in most advanced industrial countries have generally not been generating their own specialists in the area. They have mostly had to rely on a trickle of computer scientists, software engineers and analysts coming across from STEM sciences like medicine and IT. Arguably the social sciences are now converging strongly with these three and other STEM disciplines (Dunleavy et al, 2014; Bastow et al, 2014, Ch.1), so that these flows may increase. Yet business demand for 'big data' experts is soaring, and the most interesting and innovative projects are being undertaken by giant firms like Google, Amazon or Apple. So ordinary civil government departments (not intelligence and defence agencies) have major difficulties in finding talented staff (or even recruiting well-qualified consultancy firms), to undertake regular 'big data' analysis. The university sector is often in a position to help here, along with some NGOs and small and medium sized agencies. But government departments in big nations (like the UK and USA) are often reluctant to let them in to policy-making and implementation systems long-term, and do not want to become dependent upon them (as they are with many consultancies supplying staff for complex IT roles).

For the social sciences there are also risks inherent in becoming too closely engaged in applied work for policy-makers, especially in fields where the emphasis upon achieving simple predictive control rather than necessarily advancing causal understanding informed by social or economic theory. Some authors hark back to the role that the early social sciences played from the late nineteenth century through the 1940s in generating a toolkit of statistics and methods of analysis that greatly enhanced the powers of big government. For example, Robertson and Travaglia (2015) recently argued: 'We run the risk in the social sciences of perpetuating the ideological victories of the first data revolution as we progress through the second'.

Conclusions

Like many other technological and IT-related developments before it, the advent of big data has ambiguous implications (Bloom et al, 2009). On the one hand the new information-access opportunities can be decentralizing by placing into the hands of relatively junior (even grass roots or street-level) public sector staff the capacity to make far better informed decisions (von Hippel, 2006), and enhancing the competencies to deliver public services that are better attuned to citizens' needs. On the other hand, big data clearly potentially enhances the communications or network control capacities of central decision-makers, their ability to shape social behaviours and achieve timely (perhaps even real-time) effects. This latter capability may be differentially developed only by large corporations on the one hand, or in 'dark side' areas of government like security and intelligence, within a climate of weak legal and constitutional protections for civil liberties and personal privacy. Here especially there is a large potential for citizens' resistance (and guerrilla hacking) to slow big data advances across civil government.

Yet if development of 'big data' competencies could be broadened via accessible university research, and if the multiple tensions around data security and privacy issues could be better handled in future, then this somewhat dystopian future is not automatic. There remains a considerable potential for fruitful and balanced advances in social knowledge, linked with innovations in theory-based social science. Big data is also an important area of potential advance in government, especially with strengthened protections, on which see Crawford and Schultz. 2014). In Anglo-American democracies especially, reform here could set the stage for the growth of more agile, expert and research-based central state policy-making; and for more sensitive, personalized, effective and timely (even preventative) delivery of public services.

Note

I am deeply indebted to Dr Michael Jensen of the Institute for Governance and Policy Analysis, University of Canberra for several detailed conversations about big data, from which I learnt a huge amount.

References

Administrative Data Research Network (UK) (2016) Website at http://adrn.ac.uk/ (accessed 4 January 2016).

Anderson, Chris. (2008). 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'. *Wired* vol.16, no. 07.

Anderson, C. (2009) *Free! The Future of a Radical Price*. Hyperion.

Anduiza, Eva; Jensen, Michael J. and Jorba, Laia (eds). (2012) *Digital Media and Political Engagement Worldwide: A Comparative Study*,. Cambridge: Cambridge University Press.

Armstrong, Harry. (2015) *Machines That Learn in the Wild: Machine learning capabilities, limitations and implications* (London: NESTA). http://bit.ly/1TDJSyo (Accessed 5 January 2016).

Bates, David W. et al. 2014. 'Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients.' *Health Affairs* 33(7): 1123–31.

Bélanger, France, and Robert E. Crossler. 2011. 'Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems.' *Management Information Systems Quarterly*. Vol. 35, no. 4, pp. 1017–42.

Bholat, David. 'Big data and central banks', (London: Bank of England, Centre for Central Banking Studies). www.bankofengland.co.uk/research/Documents/ccbs/bigdatawriteup.pdf

Bholat, David, Hansen, Stephen, Santos, Pedro and Schonhardt-Bailey, Cheryl (2015) 'Text mining for central banks: handbook'. *Centre for Central Banking Studies* (33). pp. 1-19. ISSN 1756-7270. Available at: http://eprints.lse.ac.uk/62548/ (Accessed 5 January 2016).

Bloom, Nicholas; Garicano, Luis; Sadun, Raffaella and Van Reenen, John. (2009) 'The distinct effects of Information Technology and Communication Technology on firm organization', *NBER Working Paper*, No. 14975

Boyd, Danah, and Kate Crawford. 2012. 'Critical Questions for Big Data Provocations for a cultural, technological, and scholarly phenomenon: .' *Information, Communication and Society* 15(5): 662–79.

Bradshaw, Paul (2014) *Scraping for Journalists* (London: Online Journalism Blog). Kindle edition.

Cabinet Office. (2013) *G8 Open Data Charter and Technical Annex* http://bit.ly/1lRGFxq (Accessed 5 January 2016).

Christakis, Nicholas (2013) 'Let's shake up the social sciences', *New York Times Sunday Review*, 19 July. nyti.ms/1euewoY (Accessed 6 January 2016).

Constantiou, Ioanna D., and Jannis Kallinikos. 2014. 'New Games, New Rules: Big Data and the Changing Context of Strategy.' *Journal of Information Technology*. http://bit.ly/1ZNdX28 (Accessed 13 January, 2015).

Cook, John D. (2008) 'Five criticisms of significance testing', Blogpost, 15 November. http://bit.ly/1PzRYGH (accessed 5 January, 2016).

Crawford, Kate, and Jason Schultz. 2014. 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms.' *Boston College Law Review* 55: 93. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2325784

Cukier, Kenneth and Mayer-Schonberger, Victor. (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*. (London: Hodder).

Devos, Ken and Zackrisson, Marcus. (2015) 'Tax compliance and the public disclosure of tax information: An Australia/Norway Comparison', *eJournal of Tax Research* (2015) vol. 13, no.1, pp. 108-129.

Dumbill, Edd. 2012. 'What Is Big Data? - O'Reilly Radar.' http://oreil.ly/1O4CGK6 (Accessed 10 May, 2012).

Dunleavy, Patrick; Bastow, Simon and Tinkler, Jane. (2014) 'The contemporary social sciences are now converging strongly with STEM disciplines in the study of "human-dominated systems" and "human-influenced systems"', LSE *Impact of the Social Sciences* blog, 20 January. http://bit.ly/1mzJodJ (Accessed 6 January 2016).

Dunleavy, Patrick; Helen Margetts, Simon Bastow and Jane Tinkler. (2006) *Digital Era Governance: IT Corporations, the State, and e-Government* (Oxford: Oxford University Press).

Escher, Tobias; Margetts, Helen; Cox, Ingemar J. and Petricek, Vaclav. (2006)'Governing from the Centre? Comparing the Nodality of Digital Governments',(Paper to the Annual Meeting of the American Political Science Association, Philadelphia, 31 August - 4 September 2006). Free download at: http://www.governmentontheweb.org/access_papers.asp#J (Accessed 4 January 2016).

European Union. (2016). Website on 'Protection of Personal Data'. Directorate-General, Justice. http://ec.europa.eu/justice/data-protection/ (Accessed 5 January 2016).

Frank, Scott et al. 2013. 'Privacy Control of Location Information.' http://www.google.com/patents/US8489110 (May 27, 2015).

Gammera, Nick; Tom Cherrettb and Christopher Gutteridgec. (2014) 'Disseminating real-time bus arrival information via QRcode tagged bus stops: a case study of user take-up and reaction in Southampton, UK', *Journal of Transport Geography*, Vol. 34, pp. 254–261.

Gates, Kelly. (2011) *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance* (New York: NYU Press).

GDS. (2014) 'Anomaly detection: A machine-learning approach', (London: Government Data Service), blog. 15 August. http://bit.ly/1oVsmcp

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. and Brilliant, L. (2009). 'Detecting influenza epidemics using search engine query data'. *Nature*, 457, pp. 1012–1014.

Gleick, James. (1987) *Chaos Theory: Making a New Science* (New York: Viking).

Government Office for Science (2012) 'Foresight. Regulatory scrutiny of algorithmic trading systems: an assessment of the feasibility and potential economic impact.' Economic Impact Assessment EIA16. http://bit.ly/1RgV5Yk (Accessed 4 January 2016).

Guardian (2012). 'Riot rumours on social media left police on back foot', 2 July. http://bit.ly/1O9SRUt (accessed 5 January 2016).

Guardian. (2016) 'The NSA files', website. http://www.theguardian.com/us-news/the-nsa-files (accessed 5 January 2016).

Hale, S. 2012. 'Government on the Web: Using Crawlers and Web Archives to Map Government Presence'. Paper to Conference on 'Internet, Politics, Policy 2012: Big Data, Big Challenges?', September 21–22, University of Oxford.

Hale, S. John, P. Margetts, H. and Yasseri, T. (2014) 'Investigating Political Participation and Social Information Using Big Data and a Natural Experiment', Paper to the 2014 Annual Meeting of the American Political Science Association, August 28-31, 2014.

Halpern, David. (2015) *Inside the Nudge Unit: How small changes can make a big difference* (London: W.H. Allen).

Hood, Christopher C. and Helen Z. Margetts (2007) *The Tools of Government in the Digital Age* (Basingstoke: Palgrave Macmillan). 2nd edition.

Houghton, John. 2011. Costs and Benefits of Data Provision. Centre for Strategic Economic Studies: Victoria University. http://ands.org.au/__data/assets/pdf_file/0004/394285/houghton-cost-benefit-study.pdf

Huffington Post. (2015) 'Tim Cook Says Apple Will Resist The New UK Spying Law',11 November 2015. http://huff.to/1OLBXRE (Accessed 5 January 2016).

Huffington Post (Australia) (2015) 'Obama Asks For Stricter Laws On Data Hacking And Privacy', 13 January. http://huff.to/1kNL1aF (Accessed 6 January 2016)

Ioannidis JPA (2005) 'Why most published research findings are false'. *PLoS Medicine*, vol. 2 (no.8): e124, pp. 0696-0701. bit.ly/1zzrD71 (Accessed 6 January, 2016).

Jensen, Michael J., Laia Jorba, and Eva Anduiza. 2012. 'Introduction.' In Eva Anduiza, Michael J. Jensen, and Laia Jorba (eds) *In Digital Media and Political Engagement Worldwide: A Comparative Study*, Cambridge: Cambridge University Press, 1–15.

Joachim, David. (2003) 'What Is Intelligence Chatter, Anyway?', *Slate*, blog post 12 September. http://slate.me/1kKz6Ku (Accessed 5 January 2016).

John, Peter et al. (2013). *Nudge, Nudge, Think, Think: Experimenting with Ways to Change Civic Behaviour* (London: Bloomsbury Academic).

Kallinikos, J., A. Aaltonen, and A. Marton. 2010. 'A Theory of Digital Objects.' *First Monday* 15(6-7).

Kallinikos, Jannis. 2006. *The Consequences of Information: Institutional Implications of Technological Change*. Cheltenham, UK: Edward Elgar Publishing.

Kitchen, Rob. 2014a. 'Big Data Should Complement Small Data, Not Replace Them.' *LSE Impact of Social Sciences blog*, 27 June 2014. http://blogs.lse.ac.uk/impactofsocialsciences/2014/06/27/series-philosophy-of-data-science-.rob-kitchin/ (Accessed January 8, 2015).

Kitchin, Rob. 2014b. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London: Sage).

Land Registry. (2016) Land Registry Linked Open Data BETA site at http://landregistry.data.gov.uk/app/hpi/

Larcom, Shaun; Ferdinand Rauch and Tim Willems. (2015).'The upside of the London Tube strikes', Centrepiece LSE, Autumn 2015, pp. 12-14. http://cep.lse.ac.uk/pubs/download/cp455.pdf (Accessed 4 January 2016).

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). 'The parable of Google Flu: traps in big data analysis'. *Science*, 343, pp. 1203–1205.

Margetts, H. (2013) 'Data, Data Everywhere: Open Data versus Big Data in the Quest for Transparency', N. Bowles and J. Hamilton (eds.) *Transparency in Politics and the Media: Accountability and Open Government*. London: IB Tauris.

McGinnis, John O. 2013. *Accelerating Democracy Transforming Governance through Technological Change*. Princeton, N.J.; Woodstock: Princeton University Press.

Moja, Lorenzo et al. 2014. 'Effectiveness of Computerized Decision Support Systems Linked to Electronic Health Records: A Systematic Review and Meta-Analysis.' *American Journal of Public Health* vol. 104, no.12): e12–22.

Newcombe, Todd. 2014. 'How Government Can Unlock Economic Benefits from Open Data.' http://bit.ly/16ttVrX (Accessed 26 May, 2015).

Perry, Walter et al. (2013) *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (Santa Monica CA: Rand Corporation).

Pang, Bo and Lee, Lillian. (2008) 'Opinion Mining and Sentiment Analysis', *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135. http://bit.ly/1WTK93u (Accessed 5 January 2016).

Police UK. (2016). 'Automatic Number Plate Recognition: How police forces and other law enforcement agencies use ANPR', http://bit.ly/1S3a3AP (accessed 5 January 2016).

Radziwill, Nicole. (2015) ‘Why the ban on p-values? Understanding sampling error is key to improving the quality of research’, LSE *Impact of the Social Sciences* blog, 12 March. http://bit.ly/1Uvxm4i (Accessed 5 January 2016).

Robertson, Hamish and Travaglia, Joanne. (2015) ‘Big data problems we face today can be traced to the social ordering practices of the 19th century’, LSE *Impact of the Social Sciences* blog, 13 October. http://bit.ly/1k9Ec3w (Accessed 6 January 2016).

Robillard, Matthew. (2009) ‘What Makes APIs Hard to Learn? Answers from Developers’, *IEEE Software*, November/December, pp. 27-34.

SAS. ‘What Is Big Data?’ http://bit.ly/1iK7sch (Accessed 8 January 2015).

Savage, Mike and Burrows, Roger. (2007) ‘The Coming Crisis of Empirical Sociology’, *Sociology*, Vol. 41, No. 5, pp. 885–899.

Savage, Mike and Burrows, Roger. (2009) ‘Some Further Reflections on the Coming Crisis of Empirical Sociology’, *Sociology*, Vol. 43, No. 4, pp. 762–72.

Siegler, M. G. 2010. ‘Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003.’ *TechCrunch*. http://tcrn.ch/1h55bnj (Accessed 8 January 2015).

Sinsky, Christine A., John W. Beasley, Greg E. Simmons, and Richard J. Baron. 2014. ‘Electronic Health Records: Design, Implementation, and Policy for Higher-Value Primary Care EHRs for Higher-Value Primary Care.’ *Annals of Internal Medicine* 160(10): 727–28.

Smith, G.; Noble, M., Antilla C., Gill, L., Zaida, A., Wright, G., Dibben C. and Barnes, H. (2004) The Value of Linked Administrative Records for Longitudinal Analysis. (London: Economic and Social Research Council). Report to the ESRC National Longitudinal Strategy Committee.

Snowden Surveillance Archive (2016). Website of Snowden documents at https://snowdenarchive.cjfe.org/greenstone/cgi-bin/library.cgi (Accessed 5 January 2016).

Sunstein, Cass R. and Thaler, Richard H. (2009) *Nudge: Improving Decisions About Health, Wealth and Happiness* (London: Penguin).

Taagepera, Rein. (2008). *Making Social Sciences More Scientific: The Need for Logical Models*. Oxford: Oxford University Press.

Varian, Hal R. 2014. ‘Big Data: New Tricks for Econometrics.’ *Journal of Economic Perspectives* 28(2): 3–28.

von Hippel, Eric (2006) *Democratizing Innovation*. Cambridge MA: The MIT Press.

Wasserstein, Ronald L. and Lazar, Nicole A. (2016) ‘The ASA’s statement on p-values: context, process and purpose’, *American Statistician*, March. http://dx.doi.org/10.0031305.2016.1154108

Wikipedia. (2016a) ‘Application Program interfaces’ https://en.wikipedia.org/wiki/Application_programming_interface (Accessed 15 February 2016)

Wikipedia, (2016b) ‘List of cognitive biases’ https://en.wikipedia.org/wiki/List_of_cognitive_biases (Accessed 5 January 2016).

Williamson, Ben. (2014) ‘The death of the theorist and the emergence of data and algorithms in digital social research, LSE *Impact of the Social Sciences* blog, 10 February. bit.ly/1dfOvMR (Accessed 6 January 2016).