Development Policy Review, 2016, 34 (1): 135–174

Big Data for Development: A Review of Promises and Challenges

Martin Hilbert*

The article uses a conceptual framework to review empirical evidence and some 180 articles related to the opportunities and threats of Big Data Analytics for international development. The advent of Big Data delivers a cost-effective prospect for improved decision-making in critical development areas such as healthcare, economic productivity and security. At the same time, the well-known caveats of the Big Data debate, such as privacy concerns and human resource scarcity, are aggravated in developing countries by long-standing structural shortages in the areas of infrastructure, economic resources and institutions. The result is a new kind of digital divide: a divide in the use of data-based knowledge to inform intelligent decision-making. The article systematically reviews several available policy options in terms of fostering opportunities and minimising risks.

Key words: Big Data, decision-making, innovation, Information and Communication Technology (ICT), digital divide, digital, international development

1 Introduction

The ability to 'cope with the uncertainty caused by the fast pace of change in the economic, institutional, and technological environment' has turned out to be the 'fundamental goal of organizational changes' in the information age (Castells, 2009: 165). In the same way, the design and the execution of any development strategy consist of a myriad of smaller and larger decisions that are plagued by uncertainty. From a theoretical standpoint, every decision is an uncertain, probabilistic[1] gamble based on some kind of prior information[2] (Tversky and Kahneman, 1981). If we improve the structure of prior information on which to base our estimates, our uncertainty will on average be reduced. The better the prior, the better the estimate, the better the decision. This is not merely an intuitive analogy, but one of the core theorems of information theory and provides the foundation for all kinds of

*Assistant Professor, University of California, Davis, California (hilbert@ucdavis.edu.);

[1] 'Models must be intrinsically probabilistic in order to specify both predictions and noise-related deviations from those predictions' (Gell-Mann and Lloyd, 1996: 49).

[2] According to mathematical definition, probabilities always require previous information on which we base our probabilistic scale from 0% to 100% of chance (Caves, 1990). In other words, every probability is a conditional probability.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

analytics (Rissanen, 2010).[3] The Big Data[4] paradigm (Nature Editorial, 2008) provides a vast variety of new kinds of priors and estimation techniques to inform all sorts of decisions. The impact on the economy has been referred to as 'the new oil' (Kolb and Kolb, 2013: 10). Its impact on the social sciences can be compared to the impact of the invention of the telescope for astronomy and the invention of the microscope for biology (providing an unprecedented level of fine-grained detail). This article discusses its impact on international development.

2 Characteristics of Big Data Analytics

From a historical perspective this latest stage of the ongoing Information and Communication Technology (ICT) evolution goes back to early mass-scale computing, such as the 1890 punched card-based US Census that processed some 15 million individual records, aimed at improving governance (Driscoll, 2012). The often-cited difference of today's Big Data in terms of *velocity, volume* and *variety* of data (Brynjolfsson and McAfee, 2014; Hurwitz et al., 2013) is due to recent exponential increases in (a) telecommunication bandwidth that connects a network of (b) centralised and decentralised data storage systems, which are processed thanks to (c) digital computational capacities.

(a) *Information flow*. Over two decades of digitisation, the world's effective capacity to exchange information through two-way telecommunication networks has grown from the information equivalent of two newspaper pages per person per day in 1986 (0.3 optimally compressed exabytes worldwide, 20% of which were digitised) to six entire newspapers two decades later in 2007 (65 exabytes worldwide, 99.9% digitised) (Hilbert and López, 2011). As a result, in an average minute of 2012, Google received around 2,000,000 search queries, Facebook users shared almost 700,000 pieces of content and Twitter users roughly 100,000 microblogs (James, 2012).[5] This growth has occurred in both developed and developing countries (Hilbert, 2014c; ITU, 2012). For example, the telecommunications capacity of large Asian countries, such as China, India and Russia, was substantially above what could economically be expected from these countries during the period between 1995 and 2005 (Hilbert, 2011b). The five leading countries in terms of Facebook users in 2013 included India, Brazil, Indonesia and Mexico (Statista, 2014), while in 2011 Kuwait and Brunei had more Twitter users per capita than the UK or US, Chile more than Canada, and Brazil more than France or Germany (Mocanu et al.,

[3] Note that we have to condition on real information (not 'mis-information') and that this theorem holds in general (a particular piece of information increases uncertainty).

[4] The term 'Big Data (Analytics)' is capitalised when it refers to the discussed phenomenon.

[5] In addition to these mainly human-generated telecommunication flows, surveillance cameras, health sensors and the 'Internet of Things' (including household appliances and cars) are adding an ever-growing chunk to ever-increasing data streams (Manyika et al., 2011).

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

2013). Unlike analogue information, digital information inherently leaves a trace that can be analysed (at the time or later on).

(b) *Information stock*. At the same time, our technological memory roughly doubles about every three years, growing from 2.5 exabytes in 1986 (1% digitised), to around 300 exabytes in 2007 (94% digitised) (Hilbert and López, 2011). Already, in 2010, it cost merely $600 to buy a hard disk capable of storing all the music in the world (Kelly, 2011). This increased memory has the capacity to store an ever larger part of the growing information flow. During 1986, using all of our technological storage devices (including paper, vinyl, tape and others), we could (hypothetically) have stored less than 1% of all the information that was communicated worldwide (including broadcasting and telecommunication). By 2007 this share had increased to 16% (Hilbert and López, 2012).

(c) *Information computation*. We are still only able to analyse a small percentage of the data that we capture and store (resulting in the oft-lamented 'information overload'). Currently, financial, credit card and healthcare providers discard around 80–90% of the data they generate (Zikopoulos et al., 2012; Manyika et al., 2011). The Big Data paradigm promises to turn an ever larger part of this 'imperfect, complex, often unstructured data into actionable information' (Letouzé, 2012: 6). This expectation is fuelled by the fact that our capacity to compute information has grown between two and three times as fast as our capacity to store and communicate information (60–80% annually versus 25–30% per year) (Hilbert and López, 2012). This allows us to manage the flood of the digital information with the dike of digital computation to make sense of the data.

The extraordinary quantitative growth of these three forms of digital capacity has given rise to five differentiated qualitative characteristics in the way data is treated.

(i) *Big Data is produced anyway*. The almost inevitable digital footprint in digital networks created a plethora of opportunities to find alternative low-cost data sources. Those byproducts of digital conduct can often be used to replace traditional data sources (like surveys) with proxy indicators that correlate with the variable of interest. The best-known instance is Google's celebrated use of the 50 million most common search terms to predict the spread of seasonal flu between 2003 and 2008 (Ginsberg et al., 2009; Lazer et al., 2014). The data source (search terms) is a digital byproduct, but has the potential to replace official statistics from disease prevention and control authorities with cheap real-time data. This also implies that most Big Data sources are not produced as a result of a specific research question, unlike most traditional data sources. As such, Big Data often requires interpretation after the event, rather than prior to it.

(ii) *Big Data replaces random sampling*. Being a digital footprint of what happens in the real world, Big Data often captures all there is (sampling n = universe N). For example, with a global penetration of over 95% (ITU,

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

2014) (including 75% access among those making $1 per day or less (Naef et al., 2014)), mobile phones became a universal data source. Mobile phone records can be used to infer socioeconomic, demographic and other behavioural traits (Raento et al., 2009). For example, it has been shown how the prediction of socioeconomic level in a geographic region can automatically be performed from mobile phone records (Frias-Martinez and Virseda, 2013). Prediction accuracy depends on a combination of variables (for example, predicting gender from mobile phone behaviour is surprisingly complex (Blumenstock et al., 2010)), but using data records like call duration or frequency, an accuracy level of around 80–85% is usually achieved (Frias-Martinez et al., 2010; Soto et al., 2011). Since the mobile phone is universal in most strata, there is no need for sampling.

(iii) *Big Data is often accessible in real time*. One of the most common real-time sources for Big Data is the incessant chatter in online social media. This source is especially important in developing countries, considering the acceptance of social networks in developing countries (see above) and the wide array of content they provide. The language content of Twitter microblogs has been used to approximate cultural identities, international migration and tourism mobility in such disparate countries as Malaysia, the Philippines, Venezuela and Indonesia (Mocanu et al., 2013), and it has been shown that the 140 character microblogs from Twitter provided important information about the spread of the 2010 Haitian cholera outbreak up to two weeks before it could be provided by official statistics (Chunara et al., 2012).

(iv) *Big Data merges different sources*. The often messy and incomplete digital footprint left behind by digital conduct can be compensated for by data redundancy from different sources, often referred to as ‘data fusion’. For example, Thomson Reuters MarketPsych Indices (TRMI) distils over 3 million news articles and 4 million social media sites every day through an extensively curated language framework (MarketPsych, 2014). It not only assesses different emotional states (such as confusion, pessimism, urgency, etc.), but also opinions (such as price forecasts, etc.) and specific topics (such as special events, etc.). As in most Big Data exercises, not one single row of data is complete (not everybody provides social media feeds). However, data redundancy makes up for this fact by the complementary treatment of different sources. In 2013, the company provided 18,864 separate indices, across 119 countries, curated since 1998, and updated on a day-by-day or even minute-by-minute basis. The result is a fine-grained, real-time assessment of the local, national or regional sentiment in terms of the development of relevant indicators such as wellbeing, happiness, contentment and security, and even fear, stress, urgency, optimism, trust or anger, among others. This provides a much more fine-grained and updated picture of the current state of development than the typical coarse-grained United Nations Human Development Index (which consists of merely four broad indicators: life-expectancy, adult literacy,

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

school enrolment ratio, and Gross Domestic Product per capita (UNDP, 2014)).

(v) *The full name of Big Data is Big Data Analytics*. The notion of Big Data goes far beyond the increasing quantity and quality of data, and focuses on analysis for intelligent decision-making. Independent from the specific peta-, exa- or zettabytes scale, the key feature of the paradigmatic change is that analytic treatment of data is systematically placed at the forefront of intelligent decision-making. The process can be seen as the natural next step in the evolution from the 'Information Age' and 'Information Societies' (in the sense of Bell, 1973; Beniger, 1986; Castells, 2009; Peres and Hilbert, 2010; ITU, 2014) to 'Knowledge Societies': building on the digital infrastructure that led to vast increases in information, the Big Data paradigm focuses on converting this digital information into knowledge that informs intelligent decisions. Returning to the example of Google's flu-trend (Ginsberg et al., 2009; Lazer et al., 2014), Google processed an impressive 450 million different mathematical models in order to identify 45 search terms that could predict flu outbreaks better than traditional models. In fact, several authors define Big Data in terms of the challenge inherent in its analysis (for example, Chen et al., 2012; Chen and Zhang, 2014). The result of an extensive literature review on Big Data definitions by de Mauro et al. (2014: 8) concluded that a consensual definition of Big Data would be that 'Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value'. The analytics aspect of Big Data is also the main obstacle to its application: according to a survey of more than 3,000 managers from over 30 industries in 108 countries, the primary obstacle for Big Data adoption was 'lack of understanding of how to use analytics to improve the business', which was cited four times more than 'concerns with data quality or ineffective data governance' (LaValle et al., 2011).

This fifth characteristic of Big Data has two main implications. First, Big Data Analytics are different from traditional statistical analysis because the quantity of data affects the choice of the analytical model. Machine-learning and data-mining methods, which enable algorithms to learn from data (Shalev-Shwartz and Ben-David, 2014), were often belittled during the 1990s and early 2000s, but have proved their worth forcefully during the 2010s on being applied to vast amounts of data. It is often the case that more sophisticated models work better for smaller datasets (given the small scope of the dataset, the pattern has to be partially coded in the model, increasing the complexity of the model), while quite simple (machine detected) models work very well for larger datasets (often even better). The classic example is that in text prediction tasks (such as Google's autocomplete search entries) large memory-based models work better with datasets of under 1 million words, while simple Naïve Bayes machine-learning algorithms perform better with datasets of between 1 million and 1,000 million words (Banko and Brill, 2001; Halevy et al., 2009). The amount of available data determines the choice of model.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Secondly, exploratory data mining and machine-learning methods are not guided by theory, and do not provide any interpretation of the results. They simply detect patterns and correlations. This is often referred to as 'the end of theory' due to Big Data (Anderson, 2008). For example, machines learned from Big Data that orange used cars have the best-kept engines, that passengers who preorder vegetarian meals usually make their flights, and that spikes in the sale of prepaid phone cards can predict the location of impending massacres in the Congo (Hardy, 2012a). The reader is kindly invited to speculate about potential theories behind these correlations, being aware that such speculations can often be wrong. For example, complementary investigations showed that investments in prepaid phone cards in the Congo were not caused by the planning of or fleeing from massacres, but that dollar-denominated prepaid cards were used as hedges against impending inflation arising from the anticipated chaos (ibid.). This example shows another important point. While plain Big Data correlation analysis does not automatically reveal the bigger picture of causational theories, more and better data does provide the potential to detect spurious confounding variables and to isolate potential causation mechanisms better than ever before (in the case of the prepaid phone cards by analysing complementary data on people movements and inflation trends).

3 Conceptual Framework BD4D

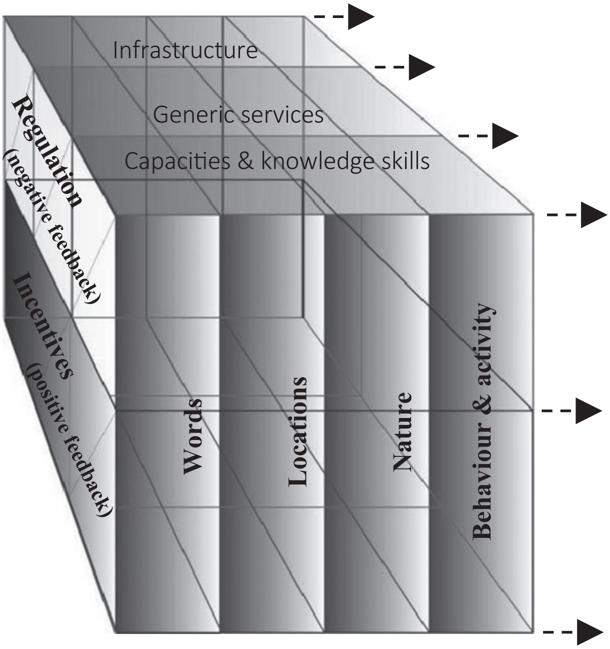
In order to be able systematically to review existing literature and related empirical evidence in the field of Big Data for Development (BD4D), we employ an established three-dimensional conceptual framework that models the process of digitisation as an interplay between technology, social change and policy strategies. The framework comes from the Information and Communication Technology for Development literature (ICT4D) (Hilbert, 2012) and is based on a Schumpeterian notion of social evolution through technological innovation (Schumpeter, 1939; Freeman and Louca, 2002; Perez, 2004). Figure 1 adapts this framework to Big Data Analytics.

The prerequisites for making Big Data Analytics work for development are a solid technological (hardware) infrastructure, generic (software) services and human capacities and skills. These horizontal layers are the essential requirements (see horizontal layers in Figure 1). They can be unequally distributed, leading to a development divide. Once available, the horizontal layers can be employed to analyse different aspects and types of data, such as words, locations, nature's elements, along with human behaviour, among others (see vertical layers in Figure 1). While this combination of technical requirements (horizontal) and social processes (vertical) is necessary for Big Data Analytics, it is not sufficient for development. The avoidance of technological determinism derives from an awareness that all technologies (including ICT) can be used to both enhance and impede capabilities (Kranzberg, 1986). Making Big Data work for development requires the social construction of its usage through carefully designed policy strategies. How can we ensure that cheap large-scale data analysis creates better public and private goods and services, rather than leading to increased State and corporate control? What needs to be considered to avoid Big Data adding to the long list of failed

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

technology transfers to developing countries? From a systems theory perspective, public and private policy choices can broadly be categorised into two groups: positive feedback (such as incentives that foster specific dynamics: putting oil into the engine) and negative feedback (such as regulations that curb particular dynamics: putting water into the engine). These are the diagonal layers of Figure 1. The result is a three-dimensional framework, whereas different circumstances (horizontal) and interventions (diagonal) intersect and effect different applications of Big Data Analytics (vertical).

Figure 1: The three-dimensional 'ICT-for development-cube' framework applied to Big Data



In the following section we review some examples of applications of Big Data for development through the tracking of words, locations, natural elements, transactions, human behaviour and economic production.[6] After illustrating some of these elements of Big Data, we look at the means in the subsequent section, specifically at the international distribution of hardware and software infrastructure and analytical skills. Last but not least, we review aspects and examples of regulatory and incentive systems to make the Big Data paradigm work for development.

[6] While the traditional IT4D cube framework uses social sectors, such as business, education, health, etc. (Hilbert, 2012), this current choice underlines the different kinds of data sources. However, this is a question of choice, not substance, Big Data practices can be analysed according to the social sectors to which they apply.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

4 Application of Big Data for Development

From a macro-perspective, it is expected that Big Data-informed decision-making will add to the existing effects of digitisation. Brynjolfsson et al. (2011) found that US firms that adopted Big Data Analytics have output and productivity that is 5–6% higher than their other investments and information technology usage would lead analysts to expect. McKinsey (Manyika et al., 2011) shows that this potential goes beyond data-intensive economic sectors, like banking, investment and manufacturing, and that several sectors with particular importance for social development are quite data intensive: education, health, government and communication host one third of the data in the US in 2010. The following section is a review of some of the micro-level examples that led to such aggregated macro-level effects of Big Data, including effects on employment, crime, water supply, mining and health.

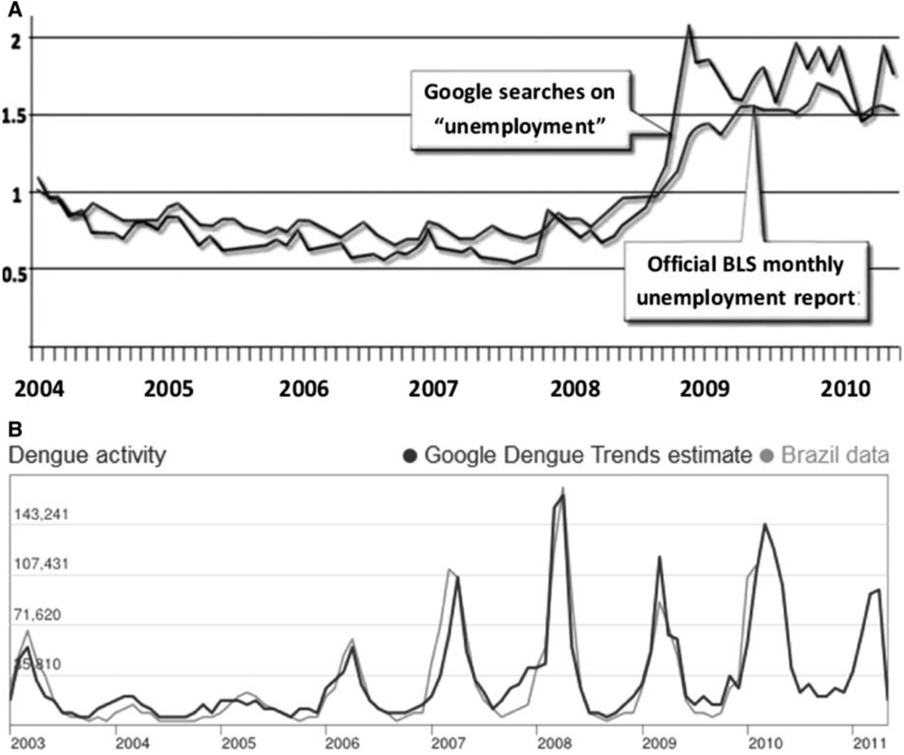
4.1 Tracking words

One of the most readily available and most structured Big Data sources relates to words. The idea is to analyse words in order to predict actions or activity. This logic is based on the old wisdom ascribed to the mystic philosopher Lao Tse: 'Watch your thoughts, they become words. Watch your words, they become actions…'. Or to say it in more modern terms: 'You Are What You Tweet' (Paul and Dredze, 2011). Figure 2a shows that the simple number of Google searches for the word 'unemployment' in the US correlates very closely with actual unemployment data from the Bureau of Labor Statistics. The latter is based on a quite expensive sample of 60,000 households and comes with a time-lag of one month, while Google trends data is available for free and in real-time (Hubbard, 2011; for a pioneering application see Ettredge et al., 2005). There are additional examples, such as search term analytics revealing trends in the Swine Flu epidemic roughly two weeks before the US Center of Disease Control (O'Reilly Radar, 2011), and similarly with dengue outbreaks (Althouse et al., 2011). Figure 2b expresses visually how Google search word trend data is able to make predictions on dengue outbreaks when official statistics from the Brazilian Ministry of Health are still unavailable.

The prototypical Big Data source for words is social media. Pioneering applications used online postings in blogs, media and web pages to predict book sales (Gruhl et al., 2005). Kalampokis et al. (2013) investigated 52 articles from 2005 to 2012 that used social media information to make social predictions (Figure 3). Social media status updates from Facebook and Twitter were the most common source, followed by blogs, review platforms and discussion boards. The applications of Big Data Analytics in this area range from swine flu pandemic (Ritterman et al., 2009) to the sales of motor vehicle parts and travel patterns (Choi and Varian, 2012). Three out of four of these studies showed clear evidence supporting the predictive and explanatory power of social media data for social phenomena (Kalampokis et al., 2013).

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Figure 2: Real-time Prediction: (a) Google searches on unemployment vs. official government statistics from the Bureau of Labor Statistics; (b) Google Brazil Dengue Activities



Source: Hubbard (2011); Google (2015).

One limitation to social media as a data source is the potential for differences between digital and real world behaviour. In a pure Goffmanian sense (Goffman, 1959), 'most of us tend to do less self-censorship and editing on Facebook than in the profiles on dating sites, or in a job interview. Others carefully curate their profile pictures to construct an image they want to project' (Manovich, 2012). The long-standing statistical issues of representativeness, biases and data-cleaning subjectivity apply to Big Data just as they do to traditional data analysis.

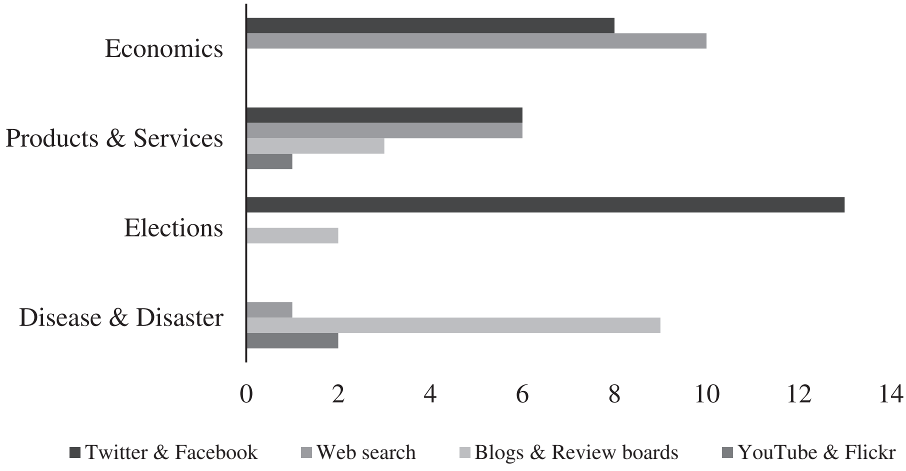
4.2 Tracking locations

The pervasiveness of mobile telephony has provided unprecedented insights into human mobility. In fact, it has been shown that the analysis of mobile phone call records enable extreme extractable predictability about human mobility, being able to predict up to 95% of people's movements in stable situations (Song et al., 2010; Lu et al., 2013) and even 85% in chaotic situations, such as after an

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

earthquake (Lu et al., 2012). Geographic mobile phone records from rural Kenya have been used to provide detailed travel and migration patterns in low-income settings to understand the spread of malaria (Buckee et al., 2013) and infectious diseases (Wesolowski et al., 2014); to understand population movements following an earthquake and cholera outbreak (Bengtsson et al., 2011; Lu et al., 2012); to study social responses to urban earthquakes in Mexico (Moumni et al., 2013); and to obtain insights into charity and reciprocal aid among peers in Rwanda after a natural disaster (Blumenstock et al., 2012). Telecom companies already sell mobility analytics obtained from mobile phones to business clients, who purchase them to gain insights into consumer behaviour in real time (Telefonica, 2012).

Figure 3: Classification of 52 Big Data social media studies by source and area of application



Source: Kalampokis et al. (2013).

While the geographical area covered by a triangulation of mobile phone base transceiver stations ranges between 1 and 3 square km (depending on whether they are in urban or rural locations), some 20–30% of mobile phones already have geo-located GSP capability, a fast growing trend (Manyika et al., 2011). Location-based services can provide much more detailed location data. In Stockholm, for example, a fleet of 2,000 GPS-equipped vehicles, consisting of taxis and trucks, provide data in 30–60 seconds intervals to create a real-time picture of the current traffic situation (Biem et al., 2010). The system can successfully predict future traffic conditions based on matching current traffic and weather data to historical records. This not only saves time and petrol, but is also useful to optimise public transport and the work of fire and police departments.

Police work and crime prediction is another important area of application (Toole et al., 2011). Chicago Crime and Crimespotting in Oakland present interactive mapping environments that allow users to track instances of crime and

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

police activity in their neighbourhood. Big Data sources such as historical crime records, along with geospatial and demographic data, can be complemented with real-time social media data, from Twitter for instance (Wang et al., 2012). Adequate algorithms and visualisation tools for developing countries are currently in the pipeline (Isafiade and Bagula, 2013).

4.3 Tracking nature

One of the biggest sources of uncertainty is nature. Reducing this uncertainty through data analysis can (i) optimise performance, (ii) mitigate risk, and (iii) improve emergency response.

(i) The attenuation from radio signals when rain falls between cellular towers has been used as a Big Data source to measure the amount of rain that falls in an area, providing crucial information to farmers and water resource managers (Overeem et al., 2013). Analysing rainfall levels, temperatures and the number of hours of sunshine, a global beverage company was able cut its beverage inventory levels by about 5% (Brown et al., 2011: 9). Relatively cheap standard statistical software was used by several bakeries to discover that the demand for cake grows with rain and the demand for salty goods with temperature. Cost savings of up to 20% have been reported as a result of fine-tuning supply and demand (Christensen, 2012). This can make the difference between the survival of a small enterprise and its failure.

(ii) Remote sensing tools have been used from as early as the late 1970s to acquire statistics on crops in developing countries and to locate petrol and mineral deposits (Paul and Mascarenhas, 1981). Nowadays robotic sensors monitor water quality and supply in river and estuary ecosystems through the movement of chemical constituents and large volumes of underwater acoustic data that track the behaviour of animals (IBM News, 2009a), as with the 315-mile Hudson River in New York state (IBM News, 2007). Similarly, the provision and analysis of data from climate scientists, local governments and communities are fused to reduce the impact of natural disasters by providing decision-makers in 25 countries with better information as to where and how to build safer schools, insure farmers against drought and protect coastal cities (GFDRR, 2012). Large datasets on weather information, satellite images and moon and tidal phases have been used to place and optimise the operation of wind turbines, estimating wind flow pattern on a grid of about 10x10 meters (32x32 feet) (IBM, 2011).

(iii) During wildfires, public authorities worldwide have started to analyse smoke patterns via real time live videos and pictorial feeds from satellites, unmanned surveillance vehicles, and specialised tasks sensors (IBM News, 2009b). Similarly, in preparation for the 2014 World Cup and the 2016 Olympics, the city of Rio de Janeiro created

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

a high-resolution weather forecasting and hydrological modelling system which gives city officials the ability to predict floods and mudslides. It has improved emergency response time by 30% (IBMSocialMedia, 2012).

4.4 Tracking transactions

Digital transactions are omnipresent footprints of social interaction (Helbing and Balietti, 2010). Analytics of sales transactions are among the most the most pervasive Big Data applications (Gruhl et al., 2005; Mayer-Schönberger and Cukier, 2013). This can go beyond the maximisation of commercial profit. Grocery and over-the-counter medication sales have been used to detect a large-scale but localised terrorism attack, such the US anthrax attacks of the early 2000s (Goldenberg et al., 2002). Besides, given that some 95% of mobile phones in developing countries are prepaid (Naef et al., 2014) and given that people put economic priority on recharging their phone, even under economic constraints (Hilbert, 2010), tracking the level of mobile phone recharging can provide a cheap source to measure poverty levels in real time on a fine-grained geographic level (Letouzé, 2012).

Historical transaction data can also be used to confront systematic abuse of social conventions. Half a century of game theory has shown that social defectors are among the most disastrous drivers of social inefficiency. A costly and often inefficient overhead is traditionally added to social transactions in order to mitigate the risk of defectors. Game theory also teaches us that social systems with memory of the past and predictive power of future behaviour can circumvent such inefficiency (Axelrod, 1984). Big Data can provide such memory and is already used to provide short-term payday loans that are up to 50% cheaper than the industry average. Default risk is judged via Big Data sources like mobile phone bills and the click-stream generated while applicants read the loan application website (Hardy, 2012b).

4.5 Tracking behaviour

Given the flood of behavioural Big Data, it is easy to define a whole new range of 'abnormal behaviour' (defined by variances around the 'average collective behaviour'). As an example from the health sector, Dartmouth (2012) presents the hospitalisation rates for forearm and hip fractures across the US. While standard deviations of hip fractures are within expected ranges, forearm fracture hospitalisation rates are 9 times higher (30% of regions with extreme values). Complementary investigations point to four general types of variation:

(i) Environmental conditions: variations in Medicare spending are not reduced when adjusting for differences in illness patterns, demographics (age, sex, race) and regional prices.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

(ii) Medical errors: some regions systematically neglect preventive measures, and others have an above average rate of mistakes.
(iii) Biased judgment: the need for costly surgery is often unclear, and systematic decision-making biases are common (Wennberg et al., 2007).
(iv) Overuse and oversupply: The number of prescribed procedures does not correlate with health outcomes, but with resource availability to prescribe procedures: more healthcare spending does not correlate with mortality ($R^2 = 0.01$), nor with underuse of preventive measures ($R^2 = 0.01$), but does correlate with additional days in hospital ($R^2 = 0.28$); more surgeries during last 6 years of life ($R^2 = 0.35$); and more visits to medical specialists ($R^2 = 0.46$), and with the availability of ten or more physicians ($R^2 = 0.43$) (Darthmouth, 2012).

With Big Data, a simple analysis of variations allows the detection of 'unwarranted variations' like (ii-iv), which originate with the underuse, overuse or misuse of medical care (Wennberg, 2011).

Behavioural data can also be produced by digital applications. Examples of behavioural data generating solutions are online games like World of Warcraft (11 million players in 2011) and FarmVille (65 million users in 2011). Students of the data produced by multi-player online games cannot only predict who is likely to leave the game and why (Borbora et al., 2011), but also psychological well-being (Shen and Williams, 2011) and educational outcomes (Ritterfeld et al., 2009). Video games are not only used to track behaviour, but also to influence it. Health insurance companies developed multiplayer online games to increase their clients' fitness levels. Such games are fed with data from insurance claims and medical records, and combine it with real-time behavioural data from the virtual world (Petrovay, 2012). Health points can be earned by checking into the gym, ordering a healthy lunch or regularly taking prescribed medicine.

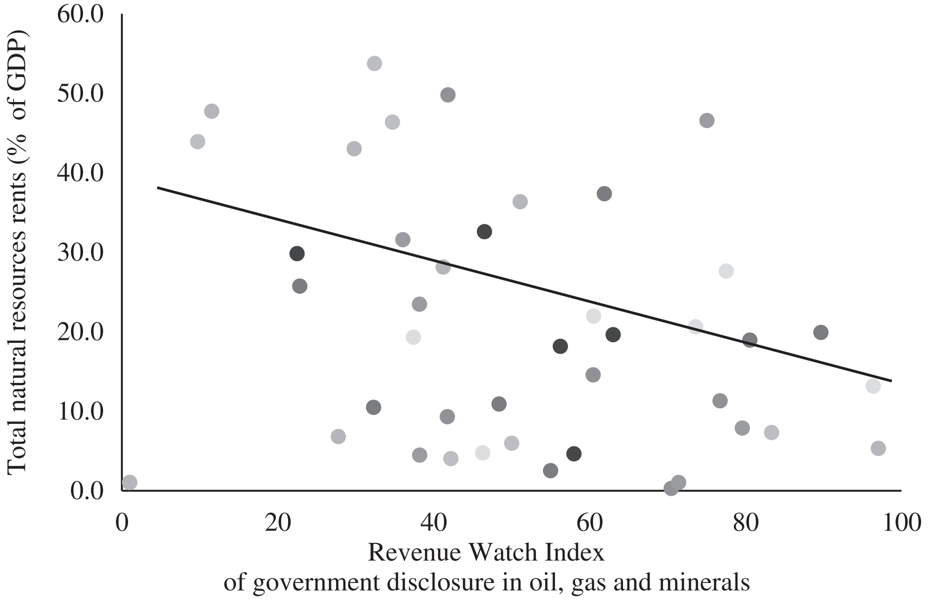
4.6 Tracking production

A contentious area of Big Data for development is the reporting of economic production that could potentially reveal competitive advantages. An illustrative case is natural resource extraction, which is a vast source of income for many developing countries (reaching from mining in South America to drilling in North Africa and the Middle East), yet have been a mixed blessing for many developing countries (often being accompanied by autocracy, corruption, property expropriation, labour rights abuses and environmental pollution). The datasets processed by resource extraction entities are enormously rich. A series of recent case studies from Brazil, China, India, Mexico, Russia, the Philippines and South Africa have argued that the publication and analysis of data that relate to the economic activity of these sectors could help to remedy the involved downsides without endangering the economic competitiveness of those sectors in developing countries (Aguilar Sánchez, 2012; Tan-Mullins, 2012; Dutta et al., 2012; Moreno, 2012; Gorre et al., 2012; Belyi and Greene, 2012; Hughes, 2012).

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

As of now, this interpretation is not in the mainstream. Figure 4 shows that the national rent that is generated from the extraction of the natural resources (revenue less cost, as percentage of GDP) negatively relates to the level of government disclosure of data on economic productivity in the oil, gas and mineral industries.

Figure 4: Public data on natural resource extraction (based on 40 countries)



Source: own elaboration, based on Revenue Watch Institute and Transparency International (2010) and World Bank (2010).

Note: The Revenue Watch Index is based on a questionnaire that evaluates whether a document, regular publication or online database provides the information demanded by the standards of the Extractive Industry Transparency Initiative (EITI), the global Publish What You Pay (PWYP) civil society movement, and the IMF's Guide on Revenue Transparency (www.revenuewatch.org/rwindex2010/methodology.html).

4.7 Tracking other data

As indicated by the right-side arrows in the conceptual framework of Figure 1, these are merely illustrative examples. Additional data sources include the tracking of financial, economic or natural resources, education attendance and grades, waste and exhaust, expenditures and investments, among many others. Future ambitions diverge as to what and how much to measure. Hardy (2012c: 4) reports of a data professional who assures that 'for sure, we want the correct name and location of every gas station on the globe … not the price changes at every station'; while his colleague interjects: 'Wait a minute, I'd like to know every gallon of gasoline that flows around the world … That might take us 20 years, but it would be interesting'.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

5 Digital Big Data Divide

Having reviewed some illustrative social ends of Big Data, let us assess the technological means (the ‘horizontal layers’ in Figure 1). The much touted digital divide (Hilbert, 2011a) is also present in the era of Big Data.

5.1 Infrastructure access

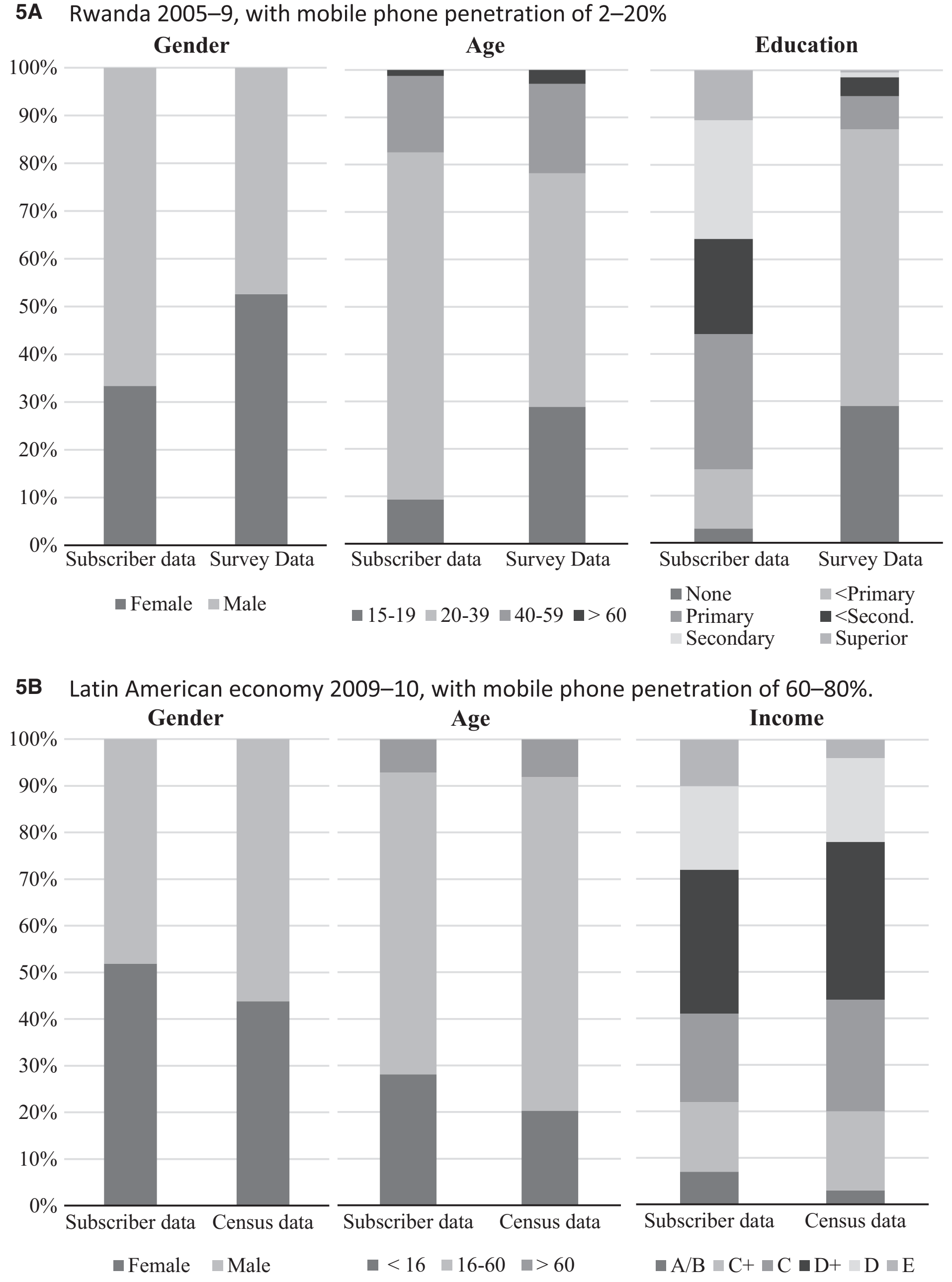
5.1.1 Challenges. ICT access inequality affects Big Data in two ways. One concerns skewed data representativeness stemming from unequal access, the other unequal access to Big Data.

(i) *Big Data is still based on samples.* While it is the ambition of Big Data to dispense with the need for random sampling techniques by collecting ‘everything there is’, students of the digital divide are very aware that the online world is only a subsample of everything there is. ‘Twitter does not represent “all people”, and it is an error to assume “people” and “Twitter users” are synonymous: they are a very particular sub-set’ (Boyd and Crawford, 2012: 669). The intensity of the bias is dictated by the intensity of the digital divide. Blumenstock and Eagle (2012: 2) showed that with low penetration rates (such as in Rwanda in 2005–2009, a period during which mobile phone penetration rose from merely 2% to 20%), ‘phone users are disproportionately male, better educated, and older’. Figure 5a shows that the mobile phone population is quite distinct from the general population, being biased toward the privileged strata. Frias-Martinez and Virseda (2013) used mobile phone data from a more advanced ‘emerging economy in Latin America’ with a mobile phone penetration of around 60–80% at the time of the study. Figure 5b shows that the Big Data sample matches official census data impressively well.

(ii) *Continuously unequal access.* Economic incentives inherent to the information economy, such as economies of scale in information storage and short product lifecycles (Shapiro and Varian, 1998), increasingly concentrate information and computational infrastructure in the ‘cloud’. While in 1986 20% of the world’s largest storage technologies were able to hold 75% of society’s technologically-stored information, this share grew to 93% by 2007. The domination of the top 20% of the world’s general-purpose computers grew from 65% in 1986, to 94% two decades later (Hilbert, 2014a). Figure 6 shows this increasing concentration of technological capacity among an ever smaller number of ever more powerful devices in the form of the Gini (1921) measure. Naturally, the vast majority of this Big Data hardware capacity resides in highly-developed countries.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Figure 5: Representativeness of Big Data: comparison of mobile phone subscribers and population at large

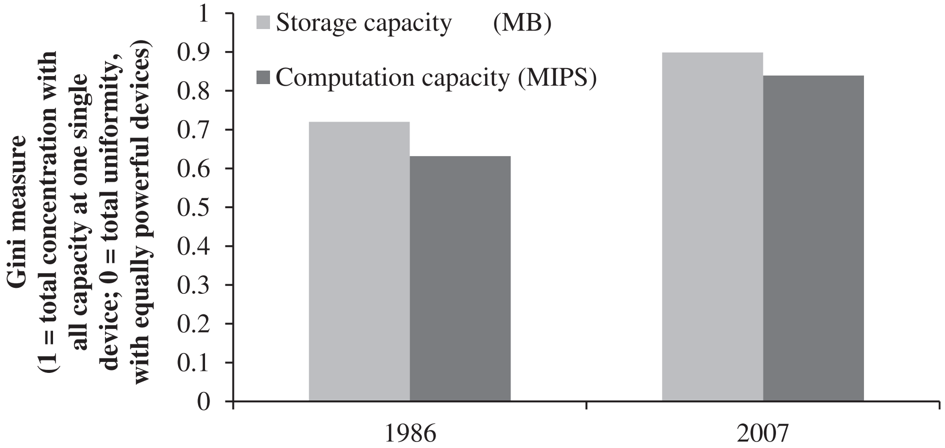


5A Rwanda 2005–9, with mobile phone penetration of 2–20%

5B Latin American economy 2009–10, with mobile phone penetration of 60–80%.

Source: (a) Blumenstock and Eagle (2012); (b) Frias-Martinez and Virseda (2013).
Note: In (b) income, the larger extreme values (segments A/B and E) are an artifact caused by the employed mapping methodology of Frias-Martinez and Virseda (2013).

Figure 6: Gini measure of the world's number of storage and computational devices, and their technological capacity (in optimally compressed MB, and MIPS), 1986 and 2007



Source: Hilbert (2014a).

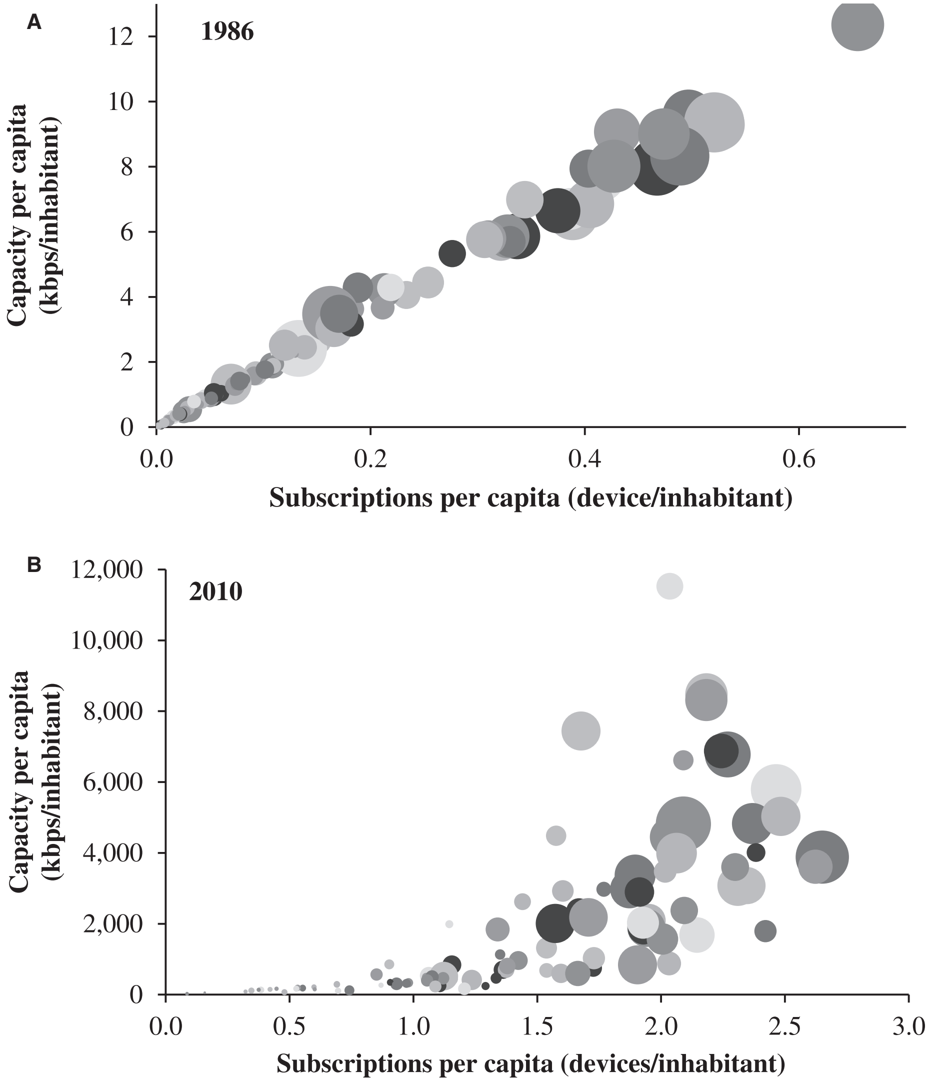
The fundamental condition to convert this increasingly concentrated information capacity among storage and computational devices ('the cloud') into an equitable information capacity among and within societies lies in the social ownership of telecommunication access. Telecommunication networks provide a potential technological gateway to the Big Data cloud. Figure 7 shows that this basic condition is further than ever from being the case. Over the past two decades, telecom access has become increasingly diversified. In the analog age of 1986, the vast majority of telecom subscriptions were fixed-line phones with similar levels of performance. This resulted in a quite linear relation between the number of subscriptions and the average traffic capacity (see Figure 7). Twenty-five years later, there are myriad different telecom subscriptions with a highly diverse performance range. Not only are telecom subscriptions heterogeneously distributed among societies, but the varied communicational performance of those channels has led to an unprecedented diversity in telecom access. Far from being closed, the digital divide incessantly evolves through an ever-changing heterogeneous collection of telecom bandwidth capacities (Hilbert, 2014c).

5.1.2 Options. One way for developing countries to confront this challenge is to create local hardware capacity by exploiting the decentralised and modular approach inherent to many Big Data solutions. Hadoop, for example, is a prominent open-source top-level Apache data-mining warehouse with a thriving community (Big Data industry leaders, such as IBM and Oracle embrace Hadoop). It is built on top of a distributed clustered file system that can take the data from thousands of distributed (also cheap low-end) PCs and server hard disks and analyse them in 64 MB blocks, which allows it to 'grow with demand while remaining economical at every size' (Shvachko et al., 2010). With respect to cost-effective distributed computational

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

power, clusters of videogame consoles are frequently used as a substitute for supercomputers in Big Data Analytics (e.g. Gardiner, 2007; Dillow, 2010). Some 500 PlayStation 3 consoles amount to the average performance of a supercomputer in 2007, which makes this option quite price competitive (López and Hilbert, 2012).

Figure 7: Subscriptions per capita vs. capacity per capita (in optimally compressed kbps of installed capacity) across 100 countries for 1986 and 2010



Note: Size of the bubbles represents Gross National Income (GNI) per capita.
Source: Hilbert (2014b).

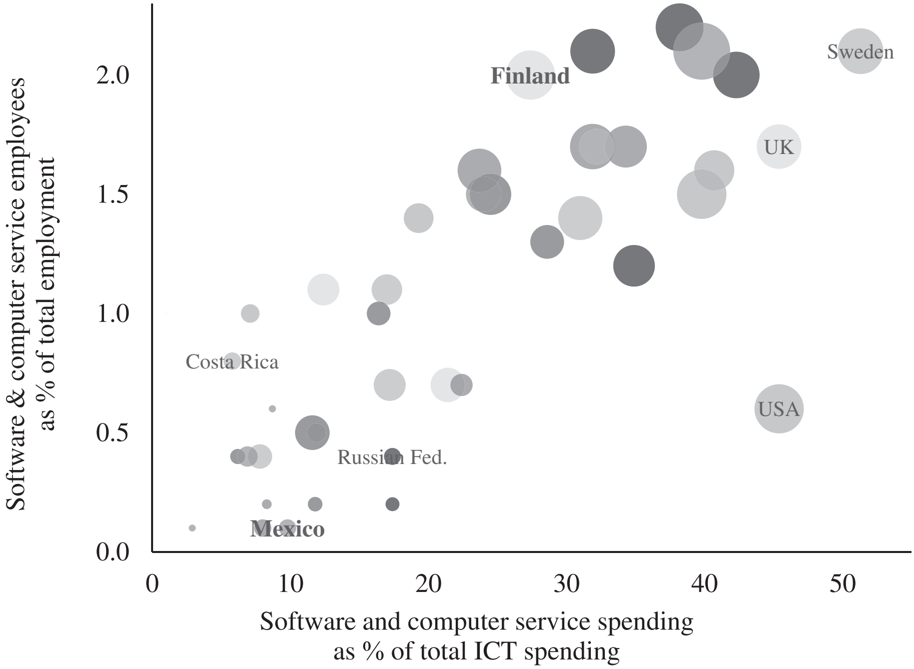
© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

5.2 Generic services

In addition to the tangible hardware infrastructure, Big Data relies heavily on software services to analyse the data. This includes both financial and human resources.

5.2.1 Challenges. Figure 8 shows the share of software and computer service spending of total ICT spending and of software and computer service employees of total employees for 42 countries. The size of the bubbles indicates total ICT spending per capita (a basic indicator for ICT advancement). Larger bubbles are related to both more software specialists and more software spending. In other words, those countries that are already behind in terms of ICT spending in absolute terms (including hardware infrastructure), have even less capacity in terms of software and computer services in relative terms. Envisioning Big Data capabilities in every enterprise, organisation and institution of a country illustrates that it makes a critical difference if 1 in 50 or 1 in 500 of the national workforce is specialised in software and computer services (see Finland as compared to Mexico in Figure 8), especially when trying to adopt and fine tune technological solutions to domestic requirements in developing countries (Romijn and Caniëls, 2011).

Figure 8: Spending and employees of software and computer services across 42 countries (as % of respective total)



Source: own elaboration, based on UNCTAD (2012).
Note: Size of bubbles represents total ICT spending per capita.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

5.2.2 Options. There are two basic options as to how to obtain Big Data services: in-house or outsourcing. Many organisations opt for a hybrid solution and use on-demand cloud resources to supplement in-house Big Data deployments (Dumbill, 2012), as in-house solutions on their own are notoriously costly (examples include large firms like Tesco, Target, Amazon and Wal-Mart). Outsourcing solutions benefit from the extremely high fixed costs and minimal variable costs of data (Shapiro and Varian, 1998): it might cost millions of dollars to create a database, but running different kinds of analysis is comparatively cheap. This economic incentive leads to an increasing agglomeration of digital data capacities in the hands of specialised data service providers (among the largest being Acxiom, Experian, Epsilon and InfoUSA). They can provide data ranging from the historic voting behaviour of politicians to evaluations of customer comments on social ratings sites like Yelp, or insights obtained from Twitter and Facebook, on-demand global trade and logistics data and information about traffic patterns and customer mobility (Hardy, 2012b, 2012c). Given their continuous flirting with the limits of the law and moral practice, they came under the scrutiny of policy-makers (US Senate, 2013). In one emblematic case, a Big Data Analytics provider classified the business attitude of millions of elderly Americans into groups like 'Elderly Opportunity Seekers: looking for ways to make money', 'Suffering Seniors: cancer or Alzheimer' and 'Oldies but Goodies: gullible, want to believe that their luck can change', and sold it to known lawbreakers who proceeded to fleece several savings accounts (Duhigg, 2007). Such obvious crimes are only the tip of the iceberg of potential discrimination due to Big Data transparency (ranging from personal credit ratings to school acceptance).

With the commoditisation of data, data also becomes subject to existing economic divides. With a global revenue of an estimated $5 billion to $10 billion in 2012–13 (Feinleib, 2012), the Big Data market has already become bigger than the size of half of the world's national economies. Creating an in-house capacity or buying the privilege of access for a fee 'produces considerable unevenness in the system: those with money – or those inside the company – can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access' (Boyd and Crawford, 2012: 673–4). In the words of fifteen leading scholars in the field: 'Computational social science is occurring – in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data … Neither scenario will serve the long-term public interest' (Lazer et al., 2009). The existing unevenness in terms of economic resources leads to an uneven playing field in this new analytic divide. Relevant policy options include financial incentives and open data policies.

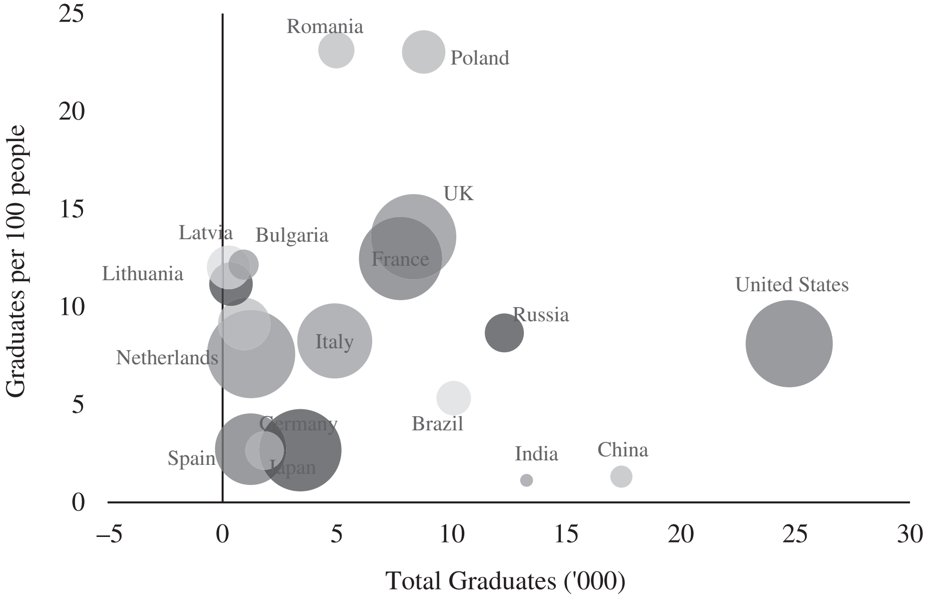
5.3 Capacities and skills

5.3.1 Challenges. Case studies on the use of Big Data applications in development show that inadequate training for data specialists and managers is one of the main

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

reasons for failure (Noormohammad et al., 2010). Manyika et al. (2011) predict that by 2018 even the United States will face a shortage of some 160,000 professionals with deep analytical skills (out of a total demand of 450,000), as well as a shortage of 1.5 million data managers capable of making informed decisions based on analytic findings (of a total of 4 million in demand). This shows that there is a global shortage, with a disproportionally negative effect on developing countries and the public sector (Borne, 2013; Freedman Consulting, 2013). Figure 9 shows that perspectives in this regard are inconsistent for different parts of the world. Some countries with relatively low income levels achieve extremely high graduation rates for professionals with deep analytical skills. In general, countries from the former Soviet bloc (such as Romania, Poland, Latvia, Lithuania and Bulgaria) produce well-trained analysts. The world's large developing BRIC economies (Brazil, Russia, India and China) produce 40% of global professionals with deep analytical skills, twice as many as the United States (far to the right on the x-axis in Figure 9). Traditional leaders of the global economy, such as Germany and Japan, are comparatively ill-equipped to satisfy domestic demand with internal sources. This has led to a long-standing and persistent discussion about brain drain and possible brain circulation in a global diaspora of top level data managers, analysts, statisticians and computer scientists.

Figure 9: Graduates with deep analytical training



Source: own elaboration, based on Manyika et al. (2011) and World Bank (2010).
Notes: i) The size of bubbles indicates countries' Gross National Income (GNI). ii) Counts people taking graduate or final-year undergraduate courses in statistics or machine learning (a subspecialty of computer science).

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

It is not only the quantity, but also the quality of analytical skills that is significant. An inventory of 52 social media studies from 2005–12 revealed that more than one third of the exercises that claimed to prove the predictive power of social media data did not even run explicit predictive analytics (using mere explanatory statistics, such as R^2 correlation analysis) (Kalampokis et al., 2013). Considering this systematic misuse of statistical techniques in the social sciences, one can only speculate about the quality of much of the Big Data research done in small enterprises, public administrations and social institutions.

5.3.2 Options. The policy implications underpinning this challenge are similar to the well-studied implications of science and technology education and the accompanying brain-drain/brain-gain from previous technological revolutions (Saxenian, 2007). One innovative way of dealing with the shortage of skilled professionals is the use of collective data analysis schemes, either through collaboration or competition. Even the most advanced users of Big Data Analytics recur to such schemes: a survey of leading scientists suggests that only a quarter of scientists have the necessary skills to analyse available data, while a third said they could obtain the skills through collaboration (Science Staff, 2011). Collaborative setups include Wikis to decode genes collectively or analyse molecular structures (Waldrop, 2008) and aid in the classification of galaxies, such as GalaxyZoo (galaxyzoo.org), and complex protein-folding problems (folding.stanford.edu). The alternative to collaboration is competition. During 2010–11 the platform Kaggle attracted over 23,000 data scientists worldwide in data analysis competitions with cash prizes of between $150 and $3,000,000 (Carpenter, 2011). While the Netflix Prize of $1,000,000 is the best known of such analytics competitions (Bell et al., 2010), a more typical example might be the 57 teams (including competitors from Chile, Serbia and Antigua and Barbuda) that helped an Australian to predict the amount of money tourists would spend in a specific area (a valuable insight for a mere $500 cash prize) (Hyndman, 2010).

6 Policy and Strategy

No technology, including Big Data, is inherently good or bad for development (Kranzberg, 1986). The maximisation of opportunities and the minimisation of risks is a process of proactive social construction, with its main societal tools being public policies and private strategies. In a development context, this starts with an awareness of the importance of data analytics and the speed with which the necessary adjustments are undertaken. The magnitude of the challenge becomes clear when considering that Ghana's statistical authorities took 17 years to adopt the UN system of national accounting proposed in 1993. In 2010 surprised statisticians found that Ghana's GDP was 62% higher than previously thought (Devarajan, 2011). It also implies considering that the ongoing transition does not occur in a vacuum, but within existing societal structures, which can result in unintended consequences. For instance, when twenty million land records in Bangalore were digitised, creating a Big Data source aimed at benefiting 7 million small farmers from over 27,000 villages (Chawla and Bhatnagar, 2004), existing elites proved much more effective at exploiting the data provided, resulting in a perpetuation of existing inequalities (Benjamin et al., 2007).

6.1 Incentives: positive feedback

One kind of intervention consists in positively encouraging and fostering desired outcomes. Two of the most common ways are providing funds for data and providing data itself.

(i) *Financial incentives and subsidies*. As so often, money is not the sole solution, but it makes things easier. Two examples from the US include the $19 billion of the American Recovery and Reinvestment Act earmarked to encourage US doctors to adopt electronic medical recordkeeping systems (Bollier, 2010), and the $700–800 million subsidies of the Office of Cyberinfrastructure (OCI) of the US National Science Foundation (NSF) for investment in 'large-scale data repositories and digitised scientific data management systems' (NSF, 2012a). Part of the desire to bring Big Data to the general public derives from investment into data visualisation (Frankel and Reid, 2008). NSF and the journal *Science* have invested in data visualisation competitions for over ten years consecutively (Norman, 2012; Science Staff, 2014).
In developing countries, small financial incentives can make a large difference. Eagle (2013) reports that a decentralised short-message-system (SMS) to collect blood bank inventory data from Kenyan hospitals initially failed because it did not consider the cost of SMS-texting (initially covered by individual nurses). By adjusting the billing system of mobile phone operators, a 1 cent subsidy was given for each SMS text message received, up to a total cost of $240 to maintain this life-saving service for 24 Kenyan hospitals.

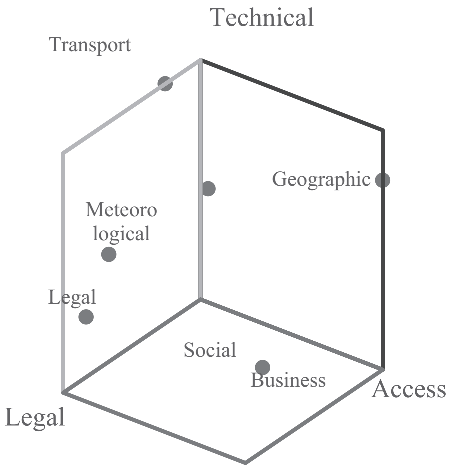
(ii) *Open data*. Another incentive for Big Data derives from providing open data to the public. This approach treats data as a public good. For example, geospatial data[7] (Vincey, 2012) and weather and climate data (GFDRR, 2012) are among the most widely published public data. The ongoing discussion about the openness of digital data moves along different dimensions (Figure 10a). It includes the use of technical standards of the data provided (such as impractical PDF files versus structured Excel spreadsheets versus machine-readable 'linked data' (Berners-Lee, 2006)), its accessibility through the web, and legal questions like the copyright and copyleft standards (Abella, 2014). The case of Spain shows that different kinds of data tend to be 'more open' with regard to one dimension or the other (Figure 10a). Data held and produced by the natural quasi-monopoly of the public sector is a natural place to push for the public provision of data, a discussion which often runs under the heading of 'open government' (Lathrop and Ruma, 2010; Kum et al., 2011; Concha and Naser, 2012; WEF and Vital Wave, 2012). Each organisation of the US government is estimated to host some 1.3 Petabytes of data, compared with a national organisational mean of 0.7 PB, while the government itself hosts around 12% of the nationally-stored data, and the related public sector areas of

[7] Geospatial data represents 37% of the datasets of the US open data initiative (Vincey, 2012).

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

education, health care and transportation another 13% (Manyika et al., 2011). In other words, if data from the public sector were to be be openly available, around a quarter of existing data resources could be liberated for Big Data Analytics. While government administrators often do not feel pressure to exploit the data they have available (Brown et al., 2011), several initiatives have pushed governments around the world to 'commit to pro-actively provide high-value information, including raw data, in a timely manner, in formats that the public can easily locate, understand and use, and in formats that facilitate reuse' (Open Government Partnership, 2014).[8] Portals like datos.gob.cl in Chile, bahrain.bh/wps/portal/data in Bahrain, or www.opendata.go.ke in Kenya provide public access to hundreds of datasets on demographics, public expenditure and natural resources. Similarly, international organisation, like the World Bank (data.worldbank.org), regional governments like Pernambuco in Brazil (dadosabertos.pe.gov.br) or local governments, like Buenos Aires in Argentina (data.buenosaires.gob.ar) provide databases about education, housing, highway conditions and the location of public bicycle stands. The potential for development is illustrated by the fact that the existence of an open data policy does not seem to correlate strongly with the level of economic well-being and perceived transparency of a country (Figure 10b). Several low-income countries are more active than their developed counterparts in making databases publicly available (for instance Kenya, Russia and Brazil), while other countries with traditionally high perceived transparency are more hesitant (such as Chile, Belgium, Sweden).

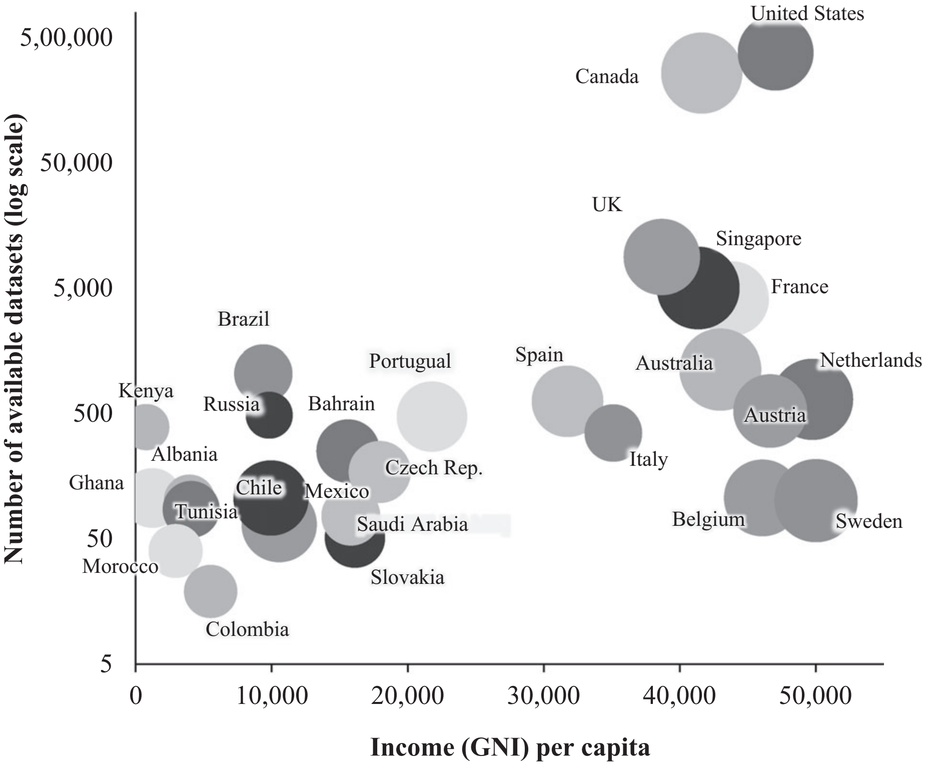
Figure 10a: Classification of type of public sector information in Spain along three open data dimensions



[8] By mid-2014, some 64 countries had signed the Open Government Declaration from which the quote is taken.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Figure 10b: Number of datasets provided on central government portal against Gross National Income per capita and Corruption Perception Index (larger data points, more transparent) for 27 countries



Source: own elaboration, based on (a) Abella, 2014; (b) 27 official open data portals; Revenue Watch Institute and Transparency International, 2011; and World Bank, 2010.
Note: The Corruption Perception Index combines the subjective estimates collected by a variety of independent institutions about the perceived level of transparency and corruption in a country (since corruption is an illegal, subjective perceptions turn out to be the most reliable method).

6.2 Regulation: negative feedback

Another way of guiding the Big Data paradigm in the desired development direction is through the establishment of regulations and legislative frameworks.

(i) *Control and privacy*. Concerns about privacy and State and corporate control through data are as old as electronic database management. Fingerprinting for the incarcerated, psychological screening for draft inductees and income tax control for working people were among the first databases to be implemented in the US before 1920 (Beniger, 1986), and

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

inspired novelists like Huxley (1932) and Orwell (1948).[9] Big Data has taken this issue to a new level. In the words of the editor of the influential computer science journal *Communications of the ACM*: 'NSA [US National Security Agency], indeed, is certainly making Orwell's surveillance technology seem rather primitive in comparison' (Vardi, 2013: 5). The fact of the matter is that digital information always leaves a potential trace that can be tracked and analysed (Andrews, 2012) and that 'any data on human subjects inevitably raise privacy issues' (Nature Editorial, 2007: 637). One common distinction concerns whether or not the tracked data is generated actively or passively, and voluntarily or involuntarily (King, 2011), leading to a 2x2 matrix. Examples include the passive but voluntary data provision to online retailers and search engines, or the passive and involuntary data provision through mobile phone locations (Andrews, 2012). Active data provision occurs via online user ratings, Facebook posts, tweets, etc.

With regard to regulation of one kind of data or other, a 2014 White House report recommended abandoning the futile attempt to regulate which data may or may not be collected: 'Policy attention should focus more on the actual uses of Big Data and less on its collection and analysis. By actual uses, we mean the specific events where something happens that can cause an adverse consequence or harm to an individual or class of individuals' (White House, 2014a: xiii). Those adverse consequences would then be severely punished (such as in the previous example of discrimination and criminal acts against the elderly (see Duhigg, 2007)). This is in line with the legal approach of not regulating the possession of something, but rather its illegal use. For example, the case of firearms in the United States follows this legal tenet. Besides foregoing practical challenges in the regulation of possession, the advantage of permitting possession consists in ensuring that potential benefits of use are not constrained. This well-known discussion about benefits, dangers and practicality is currently being held with regard to the framework for Big Data.

The currently existing legal grey zone is exemplified by the fact that several Big Data providers have opted to obtain the assurance of the customer not to abuse the provided data. For example, the company Instant Checkmate provides information on individuals drawn from criminal records, phone and address registries, professional and business licenses, voter registration, marriage records, demographic surveys and census data, etc. Before obtaining the purchased report, the consumer has to click to agree that the information will not be used to 'make decisions about consumer credit, employment, insurance, tenant screening' and not to 'spread gossip' or 'harass people whose criminal records appear on this site'. It seems difficult to control such commitments without more binding legal rules or regulation. At the same time, the self-reported data abuses of the NSA (US Government, 2012) demonstrated that even when rules and regulations are

[9] 'By comparison with that existing today, all the tyrannies of the past were half-hearted and inefficient' (Orwell, 1948; 2, 9).

in place (incomplete as they might be), the benefits of breaking those rules are so attractive that governments have admitted undertaking illegal steps to gain them. These issues are much less regulated in developing countries, whether in academia, the private sector or government agencies.

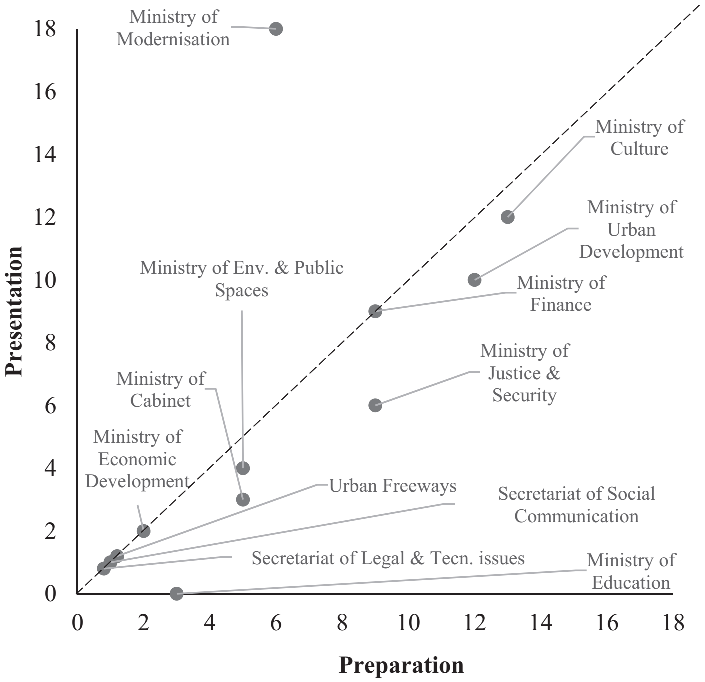
(ii) *Regulating openness*. Open government data does not need to rely on voluntary and incentivising projects, but can also be prescribed by law. So-called freedom of information legislation aims at the principle that all documents and archives of public bodies are freely accessible by each citizen, and that denial of access has to be justified by the public body and classified as an exception, not the rule. As of 2012, roughly 70 countries passed such legislation (FOI, 2012). Additionally, many developing countries have adopted transparency laws (Michener, 2009). In theory, such legislation should be the ideal catalyst for the provision of open government data. The devil in this detail is in the implementation. Table 1 shows that at least four countries in Latin America enacted such legislation, but not with any active authority or collaboration that pushes the implementation of open data forward. Whereas Costa Rica and Colombia have not enacted with such legislation, but do have concrete institutional projects. Figure 11 illustrates that the specific assignation of a leading authority can accelerate the implementation of open government directives. The Ministry of Modernisation of the city government of Buenos Aires has prepared and published the lion's share of the available databases (18 of 67), which it then provides to other authorities which present them on their webpages.

Table 1: Existence of open data legislation and implementation in Latin America

	Freedom of Information / Transparency Law	3+ entities collaborating in opening data	Governmental open data authority exists
El Salvador	Yes	Yes	Yes
México	Yes	Yes	Yes
Panama	Yes	Yes	Yes
Uruguay	Yes	Yes	Yes
Chile	Yes	Yes	No
Ecuador	Yes	No	No
Honduras	Yes	No	No
Peru	Yes	No	No
Dom.Rep.	Yes	No	No
Costa Rica	No	Yes	Yes
Colombia	No	Yes	No
E.P. Bolivia	No	No	No
R.B. Venezuela	No	No	No

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Figure 11: Number of datasets prepared and presented by government authorities in Buenos Aires



(iii) *Interoperability*. While one of the main opportunities offered by Big Data is data fusion (see above), bringing data from different sources together is also one of its main challenges. Large amounts of valuable data lurk in data silos of different departments, regional offices and specialised agencies. Manyika et al. (2011) show that the data landscape in development-relevant sectors like education and health tends to be more fragmented than those of banking or insurance services, whose databases are standardised. The regulation of data interoperability standards has become a pressing issue for the Big Data paradigm in both developed (NSF, 2012b) and developing countries (UN-ECLAC, 2007; de la Fuente, 2012).

7 Critical Reflection: all power to the algorithms?

In the past, the vast majority of information processing was executed by managers, analysts and human data crunchers (Nelson, 2008).[10] This has changed, as human

[10] In 1901, William Elkin expressed a view typical of the time when referring to 'women as measurers and computers' (Nelson, 2008: 36)

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

evaluators have been overtaken by machines in many fields. Only a decade ago we would have been surprised to have been told that artificial intelligence diagnostics outperform human aneurysm radiologists with a success rate of 95% versus 70% (Raihan, 2010). We have become accustomed to this reality quickly. When fostering this kind of approach, we inevitably give a lot of power to algorithms (Baker, 2008). By definition, algorithms can only execute processes that are programmed into them on the basis of data that has been given to them. Unfortunately, both the programming of algorithms and the nature of the data tend to have shortcomings.

First, the programmer is rarely able to consider all the intricate complexities of an environment that consists of a large number of interdependent parts that pursue different goals. While some of the results of imperfect algorithms are rather amusing (such as a book on flies that was offered for $23 million on Amazon.com by competing algorithms that forecast supply and demand patterns (Slavin, 2011), others can have disastrous consequences that affect the stability of entire economies. A well-known example is 'black-box' trading (or algorithmic trading) (Steiner, 2012). Almost non-existent in the mid-1990s, algorithmic trading was responsible for as much as 70–5% of trading volume in the US in 2009 (Hendershott et al., 2011) and has triggered several incomprehensible mass sell-offs in stock markets (triggering so-called 'flash-crashes') (Kirilenko et al., 2011, Steiner, 2012).

Second, all statistics are informed by data, which is inevitably from the past (even in the best case, the 'real-time past', it turns from present to past through the recording processes). As such, future predictions based on past data work fine as long as the future and the past follow a similar logic. If significant changes occur in the modus operandi of the system, past data does not reflect the future. Development policies aim at creating changes with a view to creating a future that is different from the past. This limits the insights obtained from Big Data (Hilbert, 2014d). To predict a future that has never been, theory-driven models are necessary. These allow variables to be adjusted with values that have never existed in statistically observable reality. As discussed in the introduction, data mining and machine-learning methods do not aim to provide such theories, they simply predict. The flood of data has made this an explicit strategy in resource-scarce developing countries like Chile: 'here we don't have the luxury to look for people who ask why things happen, we need people who detect patterns, without asking why' (Abreu, 2013). Since explaining and predicting are notoriously different (Simon, 2002; Shmueli, 2010), blind prediction algorithms can fail calamitously if the environment changes (Lazer et al., 2014), since the insights are based on the past, not on a general understanding of the overall context.

This underlines the importance of creating more flexible models that allow the exploration of theoretical scenarios that never existed before and therefore have no empirical basis. A developed Africa will not simply be a statistically extrapolated version of Europe's past development trajectory. Past data alone cannot explain what it would be like to live in a world without pollution, without hunger, without wars. A developed Africa and a sustainable world only exist 'in theory', not 'in data' (Hilbert, 2014e). The good news is that the digital age not only changes empirical data science, but also theory-driven modelling, allowing the exploration of scenarios that never existed, through computer simulation, for example. The most powerful

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

candidates are so-called agent-based models (Epstein and Axtell, 1996; Bonabeau, 2002; Gilbert and Troitzsch, 2005; Farmer and Foley, 2009). The combination of theory-driven simulation models and Big Data input to calibrate those models is becoming the new gold standard of so-called computational social science (Hilbert, 2014b). It reminds us that Big Data by itself is limited by the same constraints as all empirical work: it is exclusively post factum.

8 Conclusion

A 2014 White House report from the office of President Obama underlined that Big Data leads to 'vexing issues (big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups)', while at the same time emphasising the 'tremendous opportunities these technologies offer to improve public services, grow the economy, and improve the health and safety of our communities' (White House, 2014b). A review of over 180 pieces of mainly recent literature,[11] and several pieces of hard fact empirical evidence, has confirmed that the Big Data paradigm entails both opportunities and threats for development. On the one hand, an unprecedented amount of cost-effective data can be exploited to inform decision-making in areas that are crucial to many aspects of development, such as healthcare, security, economic productivity and disaster and resource management, among others. The extraction of actionable knowledge from the vast amount of available digital information seems to be the natural next step in the ongoing evolution from the 'Information Age' to the 'Knowledge Age'. On the other hand, the Big Data paradigm is a technological innovation and the diffusion of technological innovation is never immediate and uniform, but inescapably uneven while diffusion proceeds. This review has shown that the Big Data paradigm currently runs through an unequal diffusion process that is compromised by structural characteristics, such as the lack of infrastructure, human capital, economic resource availability and institutional frameworks in developing countries. This creates a new dimension of the digital divide: a divide in the capacity to place the analytic treatment of data at the forefront of informed decision-making and, therefore, a divide in (data-based) knowledge.

These development challenges add to the perils inherent to the Big Data paradigm, such as concerns about State and corporate control and manipulation, and the blind trust in imperfect algorithms. This shows that the advent of the Big Data paradigm is not a panacea. It is essential that this transition be accompanied and guided by proactive policy options and targeted development projects.

first submitted March 2013
final revision accepted November 2014

[11] Given the size of the phenomenon, the literature review is not exhaustive. It has been carried out during 2012–14, mainly by the identification of academic articles and reports located through Google Scholar.

References

Abella, A. (2014) 'Modelling the Economic Impact of Information Reuse in Spain'. Unpublished Master Thesis. Madrid: Universidad Rey Juan Carlos.

Abreu, R. (2013) 'Big Data and Huawei Chile'. Presented at the Complexity, Innovation and ICT, UN ECLAC, Santiago, Chile, 16 April.

Aguilar Sánchez, C. (2012) *Brazil: No Easy Miracle. Increasing transparency and accountability in the extractive industries*. Working Paper. New York, NY and London: Revenue Watch Institute and Transparency and Accountability Initiative.

Althouse, B. M., Ng, Y. Y. and Cummings, D. A. T. (2011) 'Prediction of Dengue Incidence Using Search Query Surveillance', *PLoS Negl Trop Dis* 5(8): e1258. doi:10.1371/journal.pntd.0001258.

Anderson, C. (2008) 'The End of Theory: The data deluge makes the scientific method obsolete'. *Wired Magazine*, 23 June.

Andrews, L. (2012) *I Know Who You Are and I Saw What You Did: Social networks and the death of privacy*. New York, NY: Simon and Schuster.

Axelrod, R. (1984) *The Evolution of Co-operation*. New York, NY: Basic Books.

Baker, S. (2008) *The Numerati*. Boston, MA: Houghton Mifflin Harcourt.

Banko, M. and Brill, E. (2001) 'Scaling to Very Very Large Corpora for Natural Language Disambiguation' in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.

Belbis, J. I. (2012) *Buenos Aires, Buenos Datos*. Chile: United Nations ECLAC, IDRC, OD4D.

Bell, D. (1973) *The Coming of Post-Industrial Society: A venture in social forecasting*. New York, NY: Basic Books.

Bell, R. M., Koren, Y. and Volinsky, C. (2010) 'All Together Now: A perspective on the Netflix Prize', *CHANCE* 23(1): 24–9.

Belyi, A. and Greene, S. (2012) *Russia: A complex transition. Increasing transparency and accountability in the extractive industries*. Working Paper. New York, NY and London: Revenue Watch Institute and Transparency and Accountability Initiative.

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

Beniger, J. (1986) *The Control Revolution: Technological and economic origins of the information society*. Cambridge, MA: Harvard University Press.

Benjamin, S., Bhuvaneswari, R. and Manjunatha, P. R. (2007) *Bhoomi: 'E-Governance,' or, an anti-politics machine necessary to globalize Bangalore*. CASUM-m Working Paper. Bangalore: CASUM-m.

Berners-Lee, T. (2006) 'Linked Data', 27 July. (http://www.w3.org/DesignIssues/LinkedData.html).

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Biem, A.; Bouillet E.; Feng, H.; Riabov, A.; Verscheure, O.; Rahmani, H. K. M. and Güç B. (2010) 'Real-Time Traffic Information Management Using Stream Computing'. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. http://sites.computer.org/debull/A10june/Anand.pdf.

Blumenstock, J. E., Eagle, N. and Fafchamps, M. (2012) *Risk and Reciprocity Over the Mobile Phone Network: Evidence from Rwanda*. NET Institute Working Paper. Daytona Beach, FL: NET Institute.

Blumenstock, J. E. and Eagle, N. (2012) 'Divided We Call: Disparities in access and use of mobile phones in Rwanda', *Information Technologies and International Development* 8(2): 1–16.

Blumenstock, J. E., Gillick, D. and Eagle, N. (2010) 'Who's Calling? Demographics of mobile phone use in Rwanda', presented at *AAAI Spring Symposium: Artificial Intelligence for Development*, Standford, CA, 22-4 March.

Bollier, D. (2010) *The Promise and Peril of Big Data*. Washington, DC: The Aspin Institute.

Bonabeau, E. (2002) 'Agent-based Modeling: Methods and techniques for simulating human systems', *Proceedings of the National Academy of Sciences* 99(90003): 7280–7.

Borbora, Z.; Srivastava, J.; Kuo-Wei H. and Williams, D. (2011) 'Churn Prediction in MMORPGs Using Player Motivation Theories and an Ensemble Approach', presented at *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SocialCom)*, Boston, MA, 9-11 October.

Borne, K. (2013) 'Big Data, Small World'. Presented at the TEDxGeorgeMasonU, 11 June.

Boyd, D. and Crawford, K. (2012) 'Critical Questions for Big Data', *Information, Communication and Society* 15(5): 662–79.

Brown, B., Chui, M. and Manyika, J. (2011) 'Are You Ready for the Era of "Big Data"?', *McKinsey Quarterly, McKinsey Global Institute*(October).

Brynjolfsson, E., Hitt, L. M. and Kim, H. H. (2011) 'Strength in Numbers: How does data-driven decision making affect firm performance?', *SSRN eLibrary*.

Brynjolfsson, E. and McAfee, A. (2014) *The Second Machine Age: Work, progress, and prosperity in a time of brilliant technologies*. MP3 Una edition. Brilliance Audio.

Buckee, C. O.; Wesolowski, A.; Eagle, N.; Hansen, E. and Snow, R. W. (2013) 'Mobile Phones and Malaria: Modeling human and parasite travel', *Travel Medicine and Infectious Disease* 11(1): 15–22.

Carpenter, J. (2011) 'May the Best Analyst Win', *Science* 331(6018): 698–9.

Castells, M. (2009) *The Rise of the Network Society: The information age: Economy, society, and culture Volume I* (2nd ed.). Chichester: Wiley-Blackwell.

Caves, C. (1990) 'Entropy and Information: How much information is needed to assign a probability?', in W. H. Zurek (ed.), *Complexity, Entropy and the Physics of Information*. Oxford: Westview Press.

Chawla, R. and Bhatnagar, S. (2004) 'Online Delivery of Land Titles to Rural Farmers in Karnataka, India'. Presented at the *Scaling Up Poverty Reduction: A Global Learning Process and Conference*, Shanghai, 25-7 May.

Chen, C. L. P. and Zhang, C.-Y. (2014) ‘‘Data-intensive Applications’, *Challenges, Techniques and Technologies: A survey on Big Data’, Information Sciences* 275: 314–47.

Chen, H., Chiang, R. and Storey, V. (2012) ‘Business Intelligence and Analytics: From big data to big impact’, *MIS Quarterly* 36(4): 1165–88.

Choi, H. and Varian, H. (2012) ‘Predicting the Present with Google Trends’, *Economic Record* 88(s1): 2–9.

Christensen, B. (2012) ‘Smarter Analytics: Der Bäcker und das Wetter [the baker and the weather]’. 24 April. (www.youtube.com/watch?v = dj5iWD2TVcM).

Chunara, R., Andrews, J. R. and Brownstein, J. S. (2012) ‘Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak’, *American Journal of Tropical Medicine and Hygiene* 86(1): 39–45.

Concha, G. and Naser, A. (2012) *El desafío hacia el gobierno abierto en la hora de la igualdad* (Information Society Programme No. LC/W.465). Santiago: United Nations ECLAC.

Darthmouth (2012) *Unwarranted Variations and Their Remedies: Findings from the Dartmouth Atlas of Health Care*. Hanover, NH: Dartmouth College.

De la Fuente, C. (2012) ‘Gobierno como plataforma: retos y oportunidades’, in *El desafío hacia el gobierno abierto en la hora de la igualdad*. Santiago: United Nations ECLAC.

De Mauro, A., Greco, M. and Grimaldi, M. (2014) ‘What is Big Data? A consensual definition and a review of key research topics’. Presented at the *4th International Conference on Integrated Information*, Madrid, 5-8 September.

Devarajan, S. (2011) ‘Africa’s Statistical Tragedy’, *Africa Can End Poverty Blog*. 10 June. Washington, DC: World Bank.

Dillow, C. (2010) ‘Air Force Unveils Fastest Defense Supercomputer, Made of 1,760 PlayStation 3s’, *Popsci, the Future Now*, 2 December.

Driscoll, K. (2012) ‘From Punched Cards to ‘Big Data’: A social history of database populism’, *Communication +1* 1(1): 1–33.

Duhigg, C. (2007) ‘Bilking the Elderly, With a Corporate Assist’. *The New York Times*, 20 May.

Dumbill, E. (2012) ‘What is Big Data? An introduction to the big data landscape’. *O’Reilly Radar*, 11 January.

Dutta, R., Sreedhar, R. and Ghosh, S. (2012) *India: Development at a price. Increasing transparency and accountability in the extractive industries*. Working Paper. New York, NY and London: Revenue Watch Institute and Transparency and Accountability Initiative.

Eagle, N. (2013) ‘People Are Not Cookies’. TEDxEast. (www.youtube.com/watch?v = AT2q17EhGBMandfeature = youtube_gdata_player).

Epstein, J. M. and Axtell, R. L. (1996) *Growing Artificial Societies: Social science from the bottom up*. Washington, DC: Brookings Institute.

Ettredge, M., Gerdes, J. and Karuga, G. (2005) ‘Using Web-based Search Data to Predict Macroeconomic Statistics’, *Communications of the ACM (Association for Computing Machinery)* 48(11): 87–92.

Farmer, J. D. and Foley, D. (2009) ‘The Economy Needs Agent-Based Modelling’, *Nature* 460(7256): 685–6.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Feinleib, D. (2012) *Big Data Trends*. Presented at The Big Data Group. (www.slideshare.net/bigdatalandscape/big-data-trends).

FOI (Freedom of Information) (2012) ‘Freedom of Information Legislation’, Wikipedia. Accessed 28 March.

Frankel, F. and Reid, R. (2008) ‘Big Data: Distilling meaning from data’, *Nature* 455(7209): 30.

Freedman Consulting (2013) *A Future or Failure? The flow of technology talent into government and civil society*. New York, NY and Chicago, IL: Ford Foundation and McArthur Foundation.

Freeman, C. and Louçã, F. (2002) *As Time Goes By: From the industrial revolutions to the information revolution*. Oxford: Oxford University Press.

Frias-Martinez, V. and Virseda, J. (2013) ‘Cell Phone Analytics: Scaling human behavior studies into the millions’, *Information Technologies and International Development* 9(2): 35–50.

Frias-Martinez, V., Frias-Martinez, E. and Oliver, N. (2010) ‘A Gender-centric Analysis of Calling Behaviour in a Developing Economy Using Call Detail Records’, in *AAAI 2010 Spring Symposia Artifical Intelligence for Development*, Standford, CA, 22-4 March.

Gardiner, B. (2007) ‘Astrophysicist Replaces Supercomputer with Eight PlayStation 3s’. *Wired Magazine*, 17 October.

Gell-Mann, M. and Lloyd, S. (1996) ‘Information Measures, Effective Complexity, and Total Information’, *Complexity* 2(1): 44–52.

GFDRR (Global Facility for Disaster Reduction and Recovery) (2012) ‘Open Data for Resilience Initiative (OpenDRI)’. Washington, DC: Global Facility for Disaster Reduction and Recovery, World Bank.

Gilbert, N. and Troitzsch, K. (2005) *Simulation for the Social Scientist*. Maidenhead: Open University Press.

Gini, C. (1921) ‘Measurement of Inequality of Incomes’, *Economic Journal* 31(121): 124–6.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S. and Brilliant, L. (2009) ‘Detecting Influenza Epidemics Using Search Engine Query Data’, *Nature* 457(7232): 1012–14.

Goffman, E. (1959) *The Presentation of Self in Everyday Life*. New York, NY: Anchor.

Goldenberg, A.; Shmueli, G.; Caruana, R. A. and Fienberg, S. E. (2002) ‘Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-counter Medication Sales’, *Proceedings of the National Academy of Sciences* 99(8): 5237–40.

Gorre, I., Magulgad, E. and Ramos, C. A. (2012) *Philippines: Seizing opportunities. increasing transparency and accountability in the extractive industries*. Working Paper. New York, NY and London: Revenue Watch Institute and Transparency and Accountability Initiative.

Google (2015) Google flu trends data. (www.google.org/denguetrends/about/how.html).

Gruhl, D.; Guha, R.; Kumar, R.; Novak, J. and Tomkins, A. (2005) ‘The Predictive Power of Online Chatter’, in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, IL, 21-4 August.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Halevy, A., Norvig, P. and Pereira, F. (2009) 'The Unreasonable Effectiveness of Data', *IEEE Intelligent Systems* 24(2): 8–12.

Hardy, Q. (2012a) 'Bizarre Insights from Big Data'. *New York Times Blog*, 28 March.

Hardy, Q. (2012b) 'Better Economic Forecasts, From the Cloud'. *New York Times*, 15 March.

Hardy, Q. (2012c) 'Factual's Gil Elbaz Wants to Gather the Data Universe'. *New York Times*, 24 March.

Helbing, D. and Balietti, S. (2010) 'From Social Data Mining to Forecasting Socio-Economic Crisis', *European Physical Journal* 195(1): 3–68.

Hendershott, T., Jones, C. M. and Menkveld, A. J. (2011) 'Does Algorithmic Trading Improve Liquidity?', *Journal of Finance* 66(1): 1–33.

Hilbert, M. (2014a) 'How Much of the Global Information and Communication Explosion is Driven by More, and How Much by Better Technology?', *Journal of the Association for Information Science and Technology* 65(4): 856–61.

Hilbert, M. (2014b) *ICT4ICTD: Computational social science for digital development*. ICTD2015 Singapore.

Hilbert, M. (2014c) 'Technological Information Inequality as an Incessantly Moving Target: The redistribution of information and communication capacities between 1986 and 2010', *Journal of the Association for Information Science and Technology* 65(4): 821–35.

Hilbert, M. (2014d) 'The Ultimate Limitation of Big Data for Development'. *SciDevNet*. (www.scidev.net/global/data/opinion/ultimate-limitation-big-data-development.html).

Hilbert, M. (2014e) 'Big Data Requires Big Visions for Big Change'. London: TEDxUCL. (www.youtube.com/watch?v = UXef6yfJZAI).

Hilbert, M. (2012) 'Towards a Conceptual Framework for ICT for Development: Lessons learned from the Latin American "Cube Framework"', *Information Technologies and International Development* 8(4): 243–59.

Hilbert, M. (2011a) 'The End Justifies the Definition: The manifold outlooks on the digital divide and their practical usefulness for policy-making', *Telecommunications Policy* 35(8): 715–36.

Hilbert, M. (2011b) *Mapping the Dimensions and Characteristics of the World's Technological Communication Capacity during the Period of Digitization*. Working Paper. Mauritius: International Telecommunication Union.

Hilbert, M. (2010) 'When is Cheap, Cheap Enough to Bridge the Digital Divide? Modeling income related structural challenges of technology diffusion in Latin America', *World Development* 38(5): 756–70.

Hilbert, M. and López, P. (2012) ''How to Measure the World's Technological Capacity to Communicate, Store and Compute Information?', *International Journal of Communication* 6: 956–79.

Hilbert, M. and López, P. (2011) 'The World's Technological Capacity to Store, Communicate, and Compute Information', *Science* 332(6025): 60–5.

Hubbard, D. W. (2011) *Pulse: The new science of harnessing internet buzz to track threats and opportunities*. Hoboken, NJ: John Wiley & Sons.

Hughes, T. (2012) *South Africa: A driver of change. Increasing transparency and accountability in the extractive industries*. Working Paper. New York, NY and

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

London: Revenue Watch Institute and Transparency and Accountability Initiative.

Hurwitz, J.; Nugent, A.; Halper, F. and Kaufman, M. (2013) *Big Data for Dummies*. Hoboken, NJ: Wiley.

Huxley, A. (1932) *Brave New World*. London: Chatto and Windus.

Hyndman, R. (2010) *Tourism Forecasting Part One*. (www.kaggle.com/c/tourism1).

IBM (2011) *Vestas: Turning climate into capital with big data*. Armonk, NY: IBM.

IBM News. (2009a) 'IBM Ushers In Era Of Stream Computing'. Press release, 13 May. Armonk, NY: IBM.

IBM News. (2009b) 'UMBC Researchers Use IBM Technology to Fight Rising Threats of Forest Fires'. Press release, 19 November. Armonk, NY: IBM.

IBM News (2007) 'Beacon Institute and IBM Team to Pioneer River Observatory Network'. Press release, 16 August. Armonk, NY: IBM.

IBMSocialMedia (2012) 'SmarterCities Rio: IBM helps Rio become a smarter city'. (www-03.ibm.com/press/us/en/pressrelease/33303.wss).

Isafiade, O. and Bagula, A. (2013) 'Efficient Frequent Pattern Knowledge for Crime Situation Recognition in Developing Countries', in *Proceedings of the 4th Annual Symposium on Computing for Development*, Cape Town, 6–7 December.

ITU (International Telecommunication Union) (2012) *Measuring the Information Society 2012*. Geneva: International Telecommunication Union, ITU-D.

James, J. (2012) 'Data Never Sleeps: How much data is generated every minute?'. *Domo Blog*, 8 June. (www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/).

Kalampokis, E., Tambouris, E. and Tarabanis, K. (2013) 'Understanding the Predictive Power of Social Media', *Internet Research* 23(5): 544–59.

Kelly, K. (2011). *Keynote Web 2.0 Expo SF 2011*. San Francisco, 28 March.

Kolb, J. and Kolb, J. (2013) *The Big Data Revolution*. Create Space Independent Publishing Platform

King, G. (2011) 'Ensuring the Data-Rich Future of the Social Sciences', *Science* 331 (6018): 719–21.

Kirilenko, A. A.; Kyle, A. S.; Samadi, M. and Tuzun, T. (2011) 'The Flash Crash: The impact of high frequency trading on an electronic market', *SSRN Electronic Journal*.

Kranzberg, M. (1986) 'Technology and History: "Kranzberg's Laws"', *Technology and Culture* 27(3): 544.

Kum, H.-C., Ahalt, S. and Carsey, T. M. (2011) 'Dealing with Data: Governments records', *Science* 332(6035): 1263.

Lathrop, D. and Ruma, L. (2010) *Open Government: Collaboration, transparency, and participation in practice*. Cambridge, MA: O'Reilly Media.

LaValle, S.; Lesser, E.; Shockley, R.; Hopkins, M. and Kruschwitz, N. (2010) 'Big Data, Analytics and the Path From Insights to Value', *MIT Sloan Management Review, Winter*. (http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/).

Lazer, D.; Kennedy, R.; King, G. and Vespignani, A. (2014) 'The Parable of Google Flu: Traps in big data analysis', *Science* 343(6176): 1203–5.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; . . . Van Alstyne, M. (2009) 'Computational Social Science', *Science* 323(5915): 721–3.

Letouzé, E. (2012) *Big Data for Development: Opportunities and challenges*. New York, NY: United Nations Global Pulse.

López, P. and Hilbert, M. (2012) 'Methodological and Statistical Background on the World's Technological Capacity to Store, Communicate, and Compute Information'. (www.martinhilbert.net/WorldInfoCapacity.html).

Lu, X.; Wetter, E.; Bharti, N.; Tatem, A. J. and Bengtsson, L. (2013) 'Approaching the Limit of Predictability in Human Mobility', *Scientific Reports* 3: doi:10.1038/srep02923.

Lu, X., Bengtsson, L. and Holme, P. (2012) 'Predictability of Population Displacement After the 2010 Haiti Earthquake', *Proceedings of the National Academy of Sciences* 109(29): 11576–81.

Manovich, L. (2012) 'Trending: The promises and the challenges of big social data', in M. Gold (ed.), *Debates in the Digital Humanities*. Minneapolis, MN: The University of Minnesota Press.

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C. and Hung Byers, A. (2011) *Big Data: The next frontier for innovation, competition, and productivity*. New York, NY: McKinsey and Company.

MarketPsych (2014) 'Thomson Reuters MarketPsych Indices (TRMI)'. (www.marketpsych.com/data/).

Mayer-Schönberger, V. and Cukier K. (2013) *Big Data: A revolution that will transform how we live, work, and think*. Canada: Houghton Mifflin Harcourt.

Michener, R. G. (2009) *The Surrender of Secrecy? Explaining the strength of transparency and access to information laws*. Working Paper. Rochester, NY: Social Science Research Network.

Mocanu, D.; Baronchelli, A.; Perra, N.; Gonçalves, B.; Zhang, Q. and Vespignani, A. (2013) 'The Twitter of Babel: Mapping world languages through microblogging platforms', *PLoS ONE* 8(4): e61981.

Moreno, R. (2012) *Mexico: A moment of opportunity. Increasing transparency and accountability in the extractive industries*. Working Paper. New York, NY and London: Revenue Watch Institute and Transparency and Accountability Initiative.

Moumni, B., Frias-Martinez, V. and Frias-Martinez, E. (2013) 'Characterizing Social Response to Urban Earthquakes Using Cell-phone Network Data: The 2012 Oaxaca earthquake', in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, Zurich, Switzerland, 8–12 September.

Naef, E.; Muelbert, P.; Raza, S.; Frederick, R.; Kendall, J. and Gupta, N. (2014) *Using Mobile Data for Development*. Washington, DC: and Seattle, WA: Cartesian and Bill and Melinda Gates Foundation.

Editorial, Nature (2008) 'Community Cleverness Required', *Nature* 455(7209): 1.

Nature Editorial (2007) 'A Matter of Trust', *Nature* 449(7163): 637–8.

Nelson, S. (2008) 'Big Data: The Harvard computers', *Nature* 455(7209): 36.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Noormohammad, S. F.; Mamlin, B. W.; Biondich, P. G.; McKown, B.; Kimaiyo, S. N. and Were, M. C. (2010) 'Changing Course to Make Clinical Decision Support Work in an HIV Clinic in Kenya', *International Journal of Medical Informatics* 79(3): 204–10.

Norman, C. (2012) '2011 International Science and Engineering Visualization Challenge', *Science* 335(6068): 525.

NSF (National Science Foundation) (2012a) 'About Office of Cyberinfrastructure (OCI)'. (www.nsf.gov/od/oci/about.jsp).

NSF (National Science Foundation) (2012b) 'Community-based Data Interoperability Networks (INTEROP)'. (www.nsf.gov/od/oci/about.jsp).

Open Government Partnership (2014) *Open Government Declaration*. Washington, DC: Open Government Partnership.

O' Reilly Radar (2011) *Big Data Now: Current perspectives from O'Reilly Radar*. Cambridge, MA: OReilly Media.

Orwell, G. (1948) *1984*. London: Secker & Warburg.

Overeem, A., Leijnse, H. and Uijlenhoet, R. (2013) 'Country-wide Rainfall Maps from Cellular Communication Networks', *Proceedings of the National Academy of Sciences* 110(8): 2741–5.

Paul, C. K. and Mascarenhas, A. C. (1981) 'Remote Sensing in Development', *Science* 214(4517): 139–45.

Paul, M. and Dredze, M. (2011) 'You Are What You Tweet: Analyzing Twitter for public health', in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Catalonia, Spain, 17-21 July.

Peres, W. and Hilbert, M. (2010) *Information Societies in Latin America and the Caribbean Development of Technologies and Technologies for Development*. Santiago: United Nations ECLAC.

Perez, C. (2004) 'Technological Revolutions, Paradigm Shifts and Socio-Institutional Change', in E. Reinert (ed.), *Globalization, Economic Development and Inequality: An alternative perspective*. Cheltenham: Edward Elgar.

Petrovay, N. (2012) 'Chief Technology Officer of Avivia Health (a Kaiser Permanente subsidiary)'. (www.aviviahealth.com).

Raento, M., Oulasvirta, A. and Eagle, N. (2009) 'Smartphones: An emerging tool for social scientists', *Sociological Methods and Research* 37(3): 426–54.

Raihan, I. (2010) 'Managing Big data'. Podcast. Armonk, NY: IBM.

Rissanen, J. (2010) *Information and Complexity in Statistical Modeling*. Berlin: Springer.

Ritterfeld, U.; Shen, C.; Wang, H.; Nocera, L. and Wong, W. L. (2009) 'Multimodality and Interactivity: Connecting properties of serious games with educational outcomes', *CyberPsychology and Behavior* 12(6): 691–7.

Ritterman, J., Osborne, M. and Klein, E. (2009) 'Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic'. Presented at *1st International Workshop on Mining Social Media*, Washington, DC, 14 June.

Romijn, H. A. and Caniëls, M. C. J. (2011) 'Pathways of Technological Change in Developing Countries: Review and new agenda', *Development Policy Review* 29(3): 359–80.

Saxenian, A. (2007) *The New Argonauts: Regional advantage in a global economy*. Cambridge, MA: Harvard University Press.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Schumpeter, J. (1939) *Business Cycles: A theoretical, historical, and statistical analysis of the capitalist process*. New York, NY: McGraw-Hill.

Science Staff (2014) '2013 Visualization Challenge', *Science* 343(6171): 600–10.

Science Staff (2011) 'Challenges and Opportunities', *Science* 331(6018): 692–3.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding Machine Learning: From theory to algorithms*. Cambridge: Cambridge University Press.

Shapiro, C. and Varian, H. R. (1998) *Information Rules: A strategic guide to the network economy*. Cambridge, MA: Harvard Business Press.

Shen, C. and Williams, D. (2011) 'Unpacking Time Online: Connecting internet and massively multiplayer online game use with psychosocial well-being', *Communication Research* 38(1): 123–49.

Shmueli, G. (2010) 'To Explain or to Predict?', *Statistical Science* 25(3): 289–310.

Shvachko, K.; Kuang, H.; Radia, S. and Chansler, R. (2010) 'The Hadoop Distributed File System'. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, Lake Tahoa, NV, USA, 3-7 May.

Simon, H. A. (2002) 'Science Seeks Parsimony, Not Simplicity: Searching for pattern in phenomena', in A. Zellner, H. A. Keuzenkamp and M. McAleer (eds), *Simplicity, Inference and Modelling: Keeping it sophisticatedly simple*. Cambridge: Cambridge University Press.

Slavin, K. (2011) 'How Algorithms Shape Our World'. (www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world.html).

Song, C.; Qu, Z.; Blumm, N. and Barabási, A.-L. (2010) 'Limits of Predictability in Human Mobility', *Science* 327(5968): 1018–21.

Soto, V.; Frias-Martinez, V.; Virseda, J. and Frias-Martinez, E. (2011) 'Prediction of Socioeconomic Levels Using Cell Phone Records', in J. A. Konstan, R. Conejo, J. L. Marzo and N. Oliver (eds), *User Modeling, Adaption and Personalization*. Berlin: Springer.

Statista (2014) *Statistics and Market Data on Mobile Internet and Apps*. New York, NY: Statista.

Steiner, C. (2012) *Automate This: How algorithms came to rule our world*. New York, NY: Gildan Media LLC.

Tan-Mullins, M. (2012) *China: Gradual change. Increasing transparency and accountability in the extractive industries*. Working Paper. New York, NY and London: Revenue Watch Institute and Transparency and Accountability Initiative.

Telefonica (2012) *Smart Steps*. London: Telefonica. (dynamicinsights.telefonica.com/488/smart-steps).

Toole, J. L., Eagle, N. and Plotkin, J. B. (2011) 'Spatiotemporal Correlations in Criminal Offense Records', *ACM Trans. Intell. Syst. Technol.* 2(4): 38:1–18.

Transparency International (2011) *Corruption Percpetion Index 2011*. Berlin: Transparency International.

Tversky, A. and Kahneman, D. (1981) 'The Framing of Decisions and the Psychology of Choice', *Science* 211(4481): 453–8.

UNCTAD (United Nations Conference on Trade and Development) (2012) *Information Economy Report 2012: The software industry and developing countries*. Geneva: UNCTAD.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

UNDP (United Nations Development Programme) (2014) 'Human Development Index (HDI)'. New York, NY: UNDP. (www.hdr.undp.org/en/content/human-development-index-hdi).

UN ECLAC (United Nations Economic Commission for Latin America and the Caribbean) (2007) *White Book of e-Government Interoperability for Latin America and the Caribbean Version 3.0*. Santiago: United Nations ECLAC.

US Government (2012) *Memorandum of the Chief of SID Oversight and Compliance (OC-034-12)*. Washington, DC: US Government.

Senate, U. S. (2013) *A Review of the Data Broker Industry: Collection, use, and sale of consumer data for marketing purposes*. Washington, DC: Committee on Commerce, Science, and Transportation, US Senate.

Vardi, M. Y. (2013) 'The End of the American Network', *Communications of the ACM* 56(11): 5.

Vincey, C. (2012, July) *Opendata benchmark: FR vs UK vs US*. Presented at the Dataconnexions launch conference, Google France, 10 July.

Waldrop, M. (2008) 'Big Data: Wikiomics', *Nature News* 455(7209): 22.

Wang, X., Gerber, M. S. and Brown, D. E. (2012) 'Automatic Crime Prediction Using Events Extracted from Twitter Posts', in S. J. Yang, A. M. Greenberg and M. Endsley (eds), *Social Computing, Behavioral-Cultural Modeling and Prediction*. Berlin: Springer.

WEF (World Economic Forum) and Vital Wave Consulting (2012) *Big Data, Big Impact: New possibilities for international development*. Geneva and Palo Alto, CA: WEF and Vital Wave Consulting.

Wennberg, J. E. (2011) 'Time to Tackle Unwarranted Variations in Practice', *BMJ* 342: d1513.

Wennberg, J. E.; O'Connor, A. M.; Collins, E. D. and Weinstein, J. N. (2007) 'Extending the P4P Agenda, Part 1: How Medicare can improve patient decision making and reduce unnecessary care', *Health Affairs* 26(6): 1564–74.

Wesolowski, A.; Stresman, G.; Eagle, N.; Stevenson, J.; Owaga, C.; Marube, E.; Bousema, T.; Drakeley, C.; Cox, J. and Buckee, C. O. (2014) 'Quantifying Travel Behavior for Infectious Disease Research: A comparison of data from surveys and mobile phones', *Scientific Reports* 4(5678): 1–7.

White House (2014a) *Big Data and Privacy: A technological perspective*. Washington, DC: Executive Office of the President, President's Council of Advisors on Science and Technology.

White House (2014b) *Big Data: Seizing opportunities, preserving values*. Washington, DC: Exceutive Office of the President.

Zikopoulos, P.; Eaton, C.; deRoos, D.; Deutsch, T. and Lapis, G. (2012) *Understanding Big Data: Analytics for enterprise class hadoop and streaming data*. New York, NY: McGraw-Hill.

© The Authors 2016. Development Policy Review © 2016 Overseas Development Institute.

Copyright of Development Policy Review is the property of Wiley-Blackwell and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.