Title

Big Data for Development: A Review of Promises and Challenges

Journal

Development Policy Review, 34(1)

ISSN

0950-6764

Author

Hilbert, Martin

Publication Date

2016

DOI

10.1111/dpr.12142

Peer reviewed

Big Data for Development:

A Review of Promises and Challenges

Martin Hilbert, University of California, Davis; hilbert@ucdavis.edu

Author's version Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, 34(1), 135–174. http://doi.org/10.1111/dpr.12142

Abstract

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 the cost-effective prospect to improve decision-making in critical development areas such as health care, economic productivity, and security. At the same time, all 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 data-based knowledge to inform intelligent decision-making. The article systematically reviews several available policy options to foster the opportunities and minimize the risks.

Keywords: Big Data, decision-making, innovation, ICT, digital divide, digital, international development.

Acknowledgements: The author thanks International Development Research Centre Canada (IDRC) for commissioning a more extensive study that laid the groundwork for the present article. He is also indebted to Manuel Castells, Nathan Petrovay, Francois Bar, and Peter Monge for food for thought, and to Matthew Smith, Rohan Samarajiva, Sriganesh Lokanathan, and Fernando Perini for helpful comments on draft versions, and thanks the United Nations Economic Commission for Latin America and the Caribbean (UN-CEPAL), where part of the research was undertaken. The views expressed herein are those of the author and do not necessarily reflect the views of the United Nations.

Table of Contents

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, p. 165). As such, also the design and the execution of any development strategy consist of a myriad of smaller and larger decisions that are plagued with uncertainty. From a theoretical standpoint, every decision is an uncertain, probabilistic[1] gamble based on some kind of prior information[2] (e.g. Tversky and Kahneman, 1981). If we improve the basis of prior information on which to base our estimates, our uncertainty will be reduced on average. 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 analytics (Cover and Thomas, 2006; p. 29; also 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; p.10). Its impact on the social sciences can be compared with 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.

1. 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, e.g. the 1890 punched card based U.S. 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 (McAfee and Brynjolfsson, 2012; Hurwitz, et al., 2013) is due to recent exponential increases in (a) telecommunication bandwidth that connects a network of (b) centralized and decentralized data storage systems, which are processed thanks to (c) digital computational capacities.

(a) Information flow: During the two decades of digitization, the world's effective capacity to exchange information through two-way telecommunication networks grew from the informational equivalent of 2 newspaper pages per person per day in 1986 (0.3 optimally compressed exabytes worldwide, 20 % digitized) to six entire newspapers two decades later in

[1] "Models must be intrinsically probabilistic in order to specify both predictions and noise-related deviations from those predictions" (Gell-Mann and Lloyd, 1996; p. 49).
[2] Per 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.
[3] Note that we have to condition on real information (not "miss-information") and that this theorem holds on average (a particular piece of information increases uncertainty).
[4] The term 'Big Data (Analytics)' is capitalized when it refers to the discussed phenomenon.

2007 (65 exabytes worldwide, 99.9 % digitized) (Hilbert and López, 2011; Hilbert, 2011a). 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, Ch.5). For example, the telecommunications capacity of large Asian countries, like China, India and Russia, was way above what could economically expected from these countries during the period between 1995 and 2005 (Hilbert, 2011c). 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 U.S., Chile more than Canada, and Brazil more than France or Germany (Mocanu et al. 2013). In contrary to analog information, digital information inherently leaves a trace that can be analyzed (in real-time or later on).

(b) Information stock: At the same time, our technological memory roughly doubled about every three years, growing from 2.5 exabytes in 1986 (1 % digitized), to around 300 exabytes in 2007 (94 % digitized) (Hilbert and López, 2011; 2012). Already in 2010, it cost merely US$ 600 to buy a hard disk that can store all the world's music (Kelly, 2011). This increased memory has the capacity to ever store a 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 increased to 16 % (Hilbert and López, 2012).

(c) Information computation: We are still only able to analyze a small percentage of the data that we capture and store (resulting in the often-lamented "information overload"). Currently, financial, credit card and health care 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; p. 6). This expectation is fueled by the fact that our capacity to compute information has grown two to three times as fast as our capacity to store and communicate information (60-80 % annually vs. 25-30% per year) (Hilbert and López, 2011, 2012). This allows us to fight the fire of the digital information deluge with the fire of digital computation to make sense of the data.

The quantitative explosion of these three kinds of digital capacities has led to five distinguished qualitative characteristics in the way data is treated.

[5] Additional 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).

(i) Big data is produced anyways. 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 well-known epitome is Google's illustrious use of the 50 million most common search terms to predict the spread of the 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 form 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, which is different from most traditional data sources. As such Big Data often requires post-factum, and not ex-ante interpretation.

(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, 2014) (including 75 % access among those making US$1 per day or less, Naef et al., 2014)), mobile phones became a universal data source. Mobile phone records can be used to infer socio-economic, demographic, and other behavioral trades (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; Martínez and Martínez, 2014). Prediction accuracy depends on the combination of variables (for example, predicting gender from mobile phone behavior is surprisingly tricky (Blumenstock et al., 2010; Frias-Martinez, et al. 2010)), but using data records like call duration or frequency it is generally around 80-85% (Frias-Martinez and Virseda, 2013; Blumenstock et al., 2010; 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 arrange of content they provide. The language content of Twitter microblogs has been used to approximate cultural identities, international migration and tourism mobility, including in countries like Malaysia, the Philippines, Venezuela and Indonesia (Mocanu et al. 2013), and it has been shown that the 140 character long micro blogs from Twitter contained important information about the spread of the 2010 Haitian cholera outbreak up to two weeks earlier than 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 by data redundancy from different sources, often referred to as 'data fusion'. For example, Thomson Reuters MarketPsych Indices (TRMI)

distills daily over 3 million news articles and 4 million social media sites 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 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 (e.g. not everybody provides social media feeds). However, data redundancy allows to make up for this fact by the complementary treatment of different sources. In 2013, the company provides 18,864 separate indices, across 119 countries, curated since 1998, and updated on a daily or even minute basis. The result is a fine-grained, real-time assessment of the local, national or regional sentiment in terms of development relevant indicators such as wellbeing, happiness, content, 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, school enrollment 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 increasingly 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; Masuda, 1980; 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 eventually 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 to analyze it (e.g. 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) 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" (p.8). The analytics part of big data is also the main obstacle for 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 fifths characteristic of Big Data has two main implications. For one, 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), have often been belittled during the 1990s and early 2000s, but have proven their point forcefully during the 2010s, once they were applied to vast amounts of data. It is not rarely 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 model work better with datasets under 1 million words, while simple Naïve Bayes machine learning algorithms perform better with datasets between 1 and 1,000 million words (Banko and Brill, 2001; Halevy et al. 2009). The amount of available data determines the choice of model.

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 preordered 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 turn out wrong. For example, complementary investigations showed that investments in prepaid phone cards in the Congo were not caused by the planning or fleeing of the massacre, but that dollar denominated prepaid cards were used as hedges against impending inflation arising from the anticipated chaos (Hardy, 2012a). This example shows another important point. While plain Big Data correlation analysis does not automatically reveal the bigger picture of causational theories, at the same time, more and better data provides the potential to detect spurious confounding variables and to isolate potential causation mechanisms better than ever before (e.g. in this case by analyzing complementary data on people movements and inflation trends).

2. Conceptual Framework BD4D

In order to be able to systematically 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 digitization as an interplay between technology, social change, and policy strategies. The framework comes from the ICT4D literature

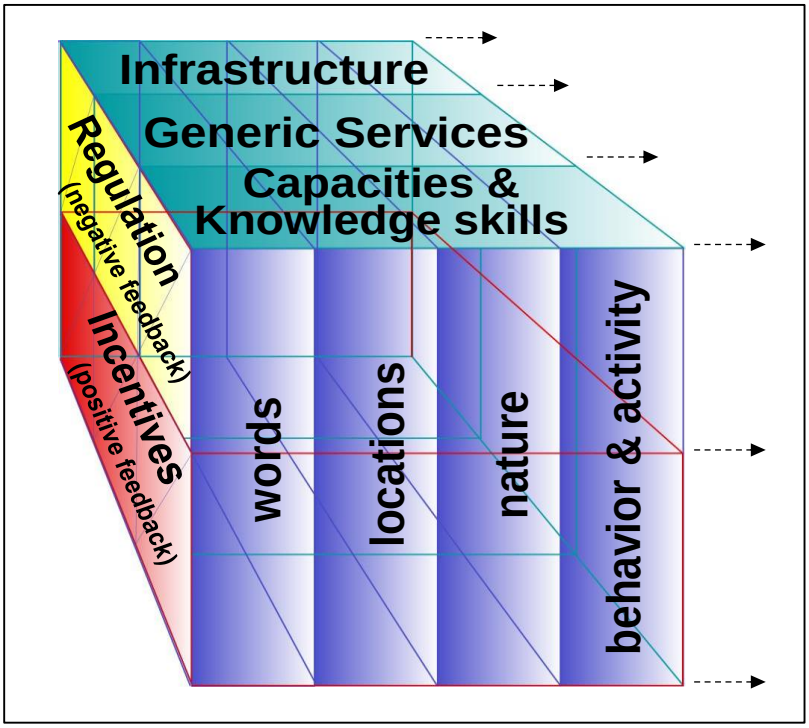
(Information and Communication Technology for Development) (Hilbert, 2012) and is based on a Schumpeterian notion of social evolution through technological innovation (Schumpeter, 1939; Freeman and Louca, 2002; Perez, 2004). Figure 1 adopts this framework to Big Data Analytics.

The first requisite for making Big Data Analytics work for development is a solid technological (hardware) infrastructure, generic (software) services, and human capacities and skills. These horizontal layers are the requirement *sine qua none* (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 analyze different aspects and kinds of data, such as words, locations, nature's elements, and human behavior, among others (see vertical layers in Figure 1). While this set-up of technical requirements (horizontal) and social processes (vertical) is necessary for Big Data Analytics, it is not sufficient for development. The rejection of technological determinism tells us that all technologies (including ICT) can be used to both foster and deprive capabilities (Kranzberg, 1986). Making Big Data work for development requires the social construction of its usage through carefully designed policy strategies. How can we assure 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 that Big Data will not add to the long list of failed technology transfer to developing countries? From a systems theoretic perspective, public and private policy choices can broadly be categorized into two groups: positive feedback (such as incentives that foster specific dynamics: putting oil into the fire), and negative feedback (such as regulations, that curb particular dynamics: putting water into the fire). These are the diagonal layers of Figure 1. The result is a three-dimensional framework, whereas different circumstances (horizontal) and interventions (diagonal) intersect and affect different applications of Big Data Analytics (vertical).

In the following section we review some examples of applications of Big Data for development through the tracking of words, locations, nature's elements, transactions, human behavior and economic production.[6] After illustrating some of these ends 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, and we might as well analyze Big Data practices according to social sectors of their application.

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



3. Application of Big Data for Development

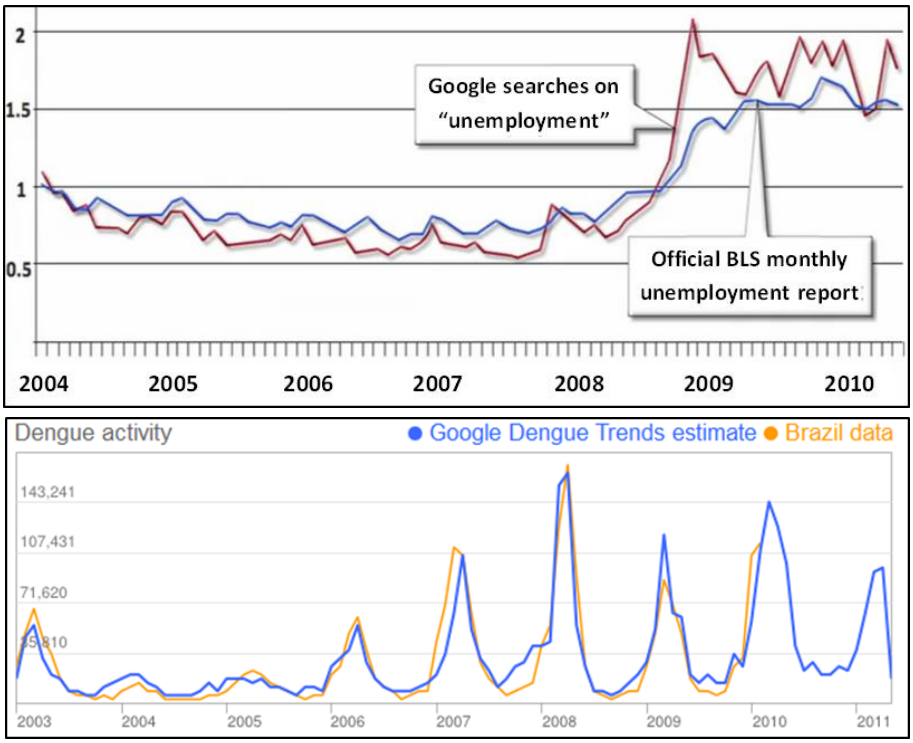
From a macro-perspective, it is expected that Big Data informed decision-making adds to the existing effects of digitization. Brynjolfsson, Hitt, and Kim (2011) found that U.S. firms that adopted Big Data Analytics have output and productivity that is 5 – 6 % higher than what would be expected given their other investments and information technology usage. 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 U.S. in 2010. The following reviews some of the micro-level examples that lead to such aggregated macro-level effects of Big Data, including effects on employment, crime, water supply, mining, and health.

3.1 Tracking words

One of the most readily available and most structured Big Data source relates to words. The idea is to analyze 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 U.S. 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 also Ettredge et al., 2005). Using a similar logic, there are several additional examples of how search term analytics was able to reveal trends in the Swine Flu epidemic roughly two weeks before the U.S. Center of Disease Control (O'Reilly Radar, 2011), and dengue outbreaks (Althouse and Ng, 2011). Figure 2b visualizes how Google search word trend data is able to make future predictions on the dengue outbreaks at times when official statistics by the Brazilian Ministry of Health are still missing.

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*

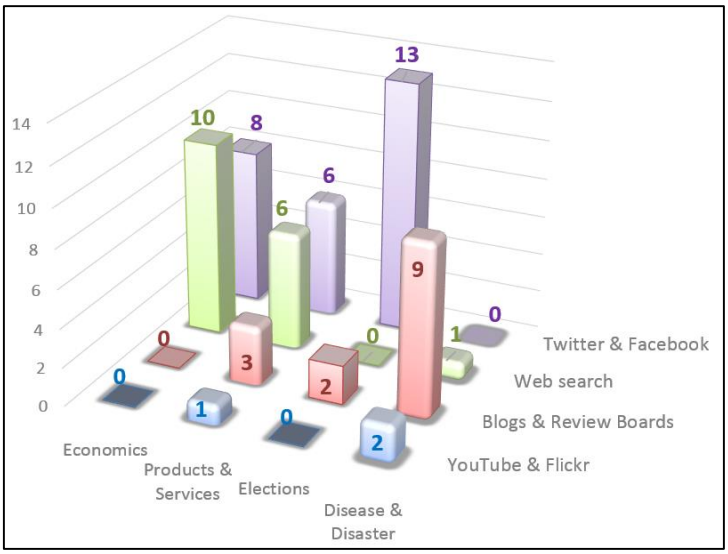


Source: Hubbard, 2011; http://www.hubbardresearch.com; Google correlate, http://www.google.org/denguetrends/about/how.html

The prototypical Big Data source for words is social media. Pioneering applications used online postings in blogs, media, and web pages to predict books 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 (see Figure 3). Social media status updates from Facebook and Twitter were the most common source, followed by blogs, review- and discussion boards. The applications of Big Data Analytics in this are 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 that supports the predictive and explanatory power of social media data for social phenomena (Kalampokis et al., 2013).

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



Source: Kalampokis, Tambouris, & Tarabanis (2013).

One limitation to social media as a data source is the potential for differences between digital and real world behavior. 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 the same as to all kinds of traditional data analysis.

3.2 Tracking locations

The above-cited 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 allow to approach the limit of extractable predictability in human mobility, being able to predict up to 95 % of people's movements in more stable situations (Song et al., 2010; Lu et al., 2013) and even 85 % in chaotic situations, such after an 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 the strike of a natural disaster

(Blumenstock et al., 2012). Telecom companies already sell the service of mobility analytics obtained from mobile phones to business clients, who purchase them to gain insights into consumer behavior in real-time (Telefonica, 2012).

While the geographical area covered by a triangulation of mobile phone base transceiver stations ranges between 1 and 3 square km (depending on their urban or rural location), 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 gasoline, but is also useful to constantly optimize public transportation, and the work of fire and police departments.

Police work and crime prediction is another important area of application (e.g. Toole, Eagle & Plotkin, 2011). Chicago Crime and Crimespotting in Oakland present interactive mapping environments that allow users to track instances of crime and police activity in their neighborhood. Big data sources such as historical crime record, geospatial and demographic data can be complemented with real-time social media data, such as from Twitter (Wang et al., 2012). Adequate algorithms and visualization tools for developing countries are currently being developed (Isafiade and Bagula, 2013).

3.3 Tracking nature

One of the biggest sources of uncertainty is nature. Reducing this uncertainty through data analysis can (i) optimize performance, (ii) mitigate risk, and (ii) 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). Analyzing 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, Chui, and Manyika, 2011, p. 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 have been used as early as in the late 1970s to acquire statistics on crops in developing countries and to locate petroleum and mineral deposits (Paul and Mascarenhas, 1981). Nowadays robotic sensors monitor water quality and supply of river and estuary

ecosystems through the movement of chemical constituents and large volumes of underwater acoustic data that tracks the behavior of animals (IBM News, May 2009), which is the case for the 315-mile New York's Hudson River (IBM News, 2007). Similarly, the provision and analysis of data from climate scientists, local governments and communities is fused to reduce the impact of natural disasters by empowering decisions-makers in 25 countries with better information on 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 optimize 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 analyze smoke patterns via real time live videos and pictorial feeds from satellite, unmanned surveillance vehicles, and specialized tasks sensors (IBM News, Nov. 2009). Similarly, in preparation for the 2014 World Cup and the 2016 Olympics, the city of Rio de Janeiro created high-resolution weather forecasting and hydrological modeling system which gives city official the ability to predict floods and mud slides. It has improved emergency response time by 30 % (IBMSocialMedia, 2012).

3.4 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 maximization of commercial profit. Grocery and over-the-counter medication sales have been used to detect a large-scale but localized terrorism attack, such the U.S. anthrax attacks of the early 2000s (Goldenberg, et al., 2002). Besides, given that some 95 % of the mobile phones in developing countries are prepaid (Naef, E. 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).

Historic 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 behavior 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's average. Default

risk is judged via Big Data sources like cellphone bills and the click-stream generated while applicants read the loan application Website (Hardy, 2012b).

3.5 Tracking behavior

Given the flood of behavioral Big Data, it is easy to define a whole new range of "abnormal behavior" (defined by variances around the "average collective behavior"). As an example from the health sector, Figure 4a presents the hospitalization rates for forearm- and hip-fractures across the U.S. (Darthmouth, 2012). While standard deviations of hip-fractures are within expected ranges, forearm fracture hospitalization rates are 9 times larger (30 % of the regions with extreme values). Complementary investigations point to four general types of variations:

(i) Environmental conditions: Figure 4b shows that variations in Medicare spending are not reduced when adjusting for differences in illness patterns, demographics (age, sex, race), and regional prices.

(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).

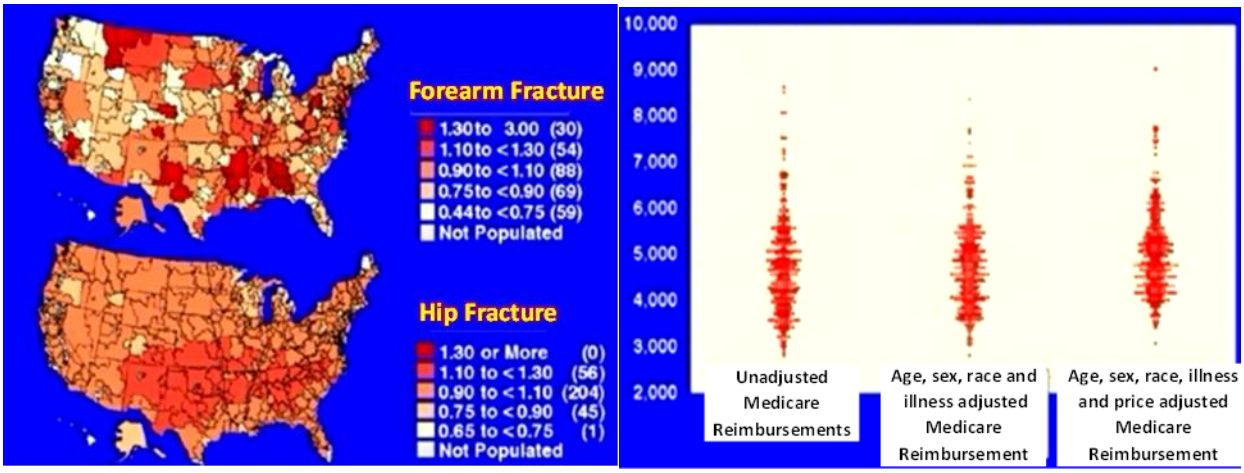
(iv) Overuse and oversupply: The number of prescribed procedures does not correlate with health outcomes, but with resource availability to prescribe procedures: more health care spending does not correlate with mortality (R^2 = 0.01), nor with underuse of preventive measures (R^2 = 0.01), but does correlate 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 to detect "unwarranted variations" like (ii-iv), which originate with the underuse, overuse, or misuse of medical care (Wennberg, 2011).

Behavioral data can also be produced by digital applications. Examples of behavioral 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, Srivastava, Hsu and Williams, 2012), but also psychological well-being (Shen and Williams, 2011) and educational outcomes (Ritterfeld et al., 2009). Video games are not only used to track, but also to influence behavior. Health insurance companies developed multiplayer online games to increase their client's fitness. Such games are fed with data from insurance claims and medical records, and combine it with real-time behavioral 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.

Figure 4: (a) *Patterns of variations in the hospitalization for forearm and hip-fracture across U.S.;* ***(b)*** *Patterns of Medicare Spending among regions in the U.S.*

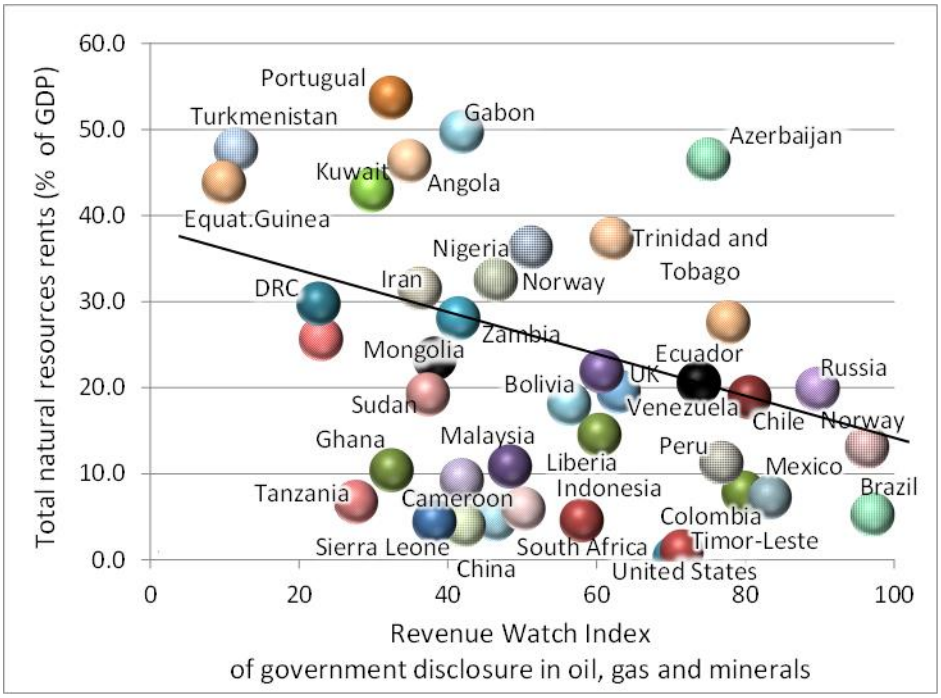


Source: Darthmouth, 2012; http://www.dartmouthatlas.org

3.6 Tracking production

A contentious area of Big Data for development is the reporting of economic production that could potentially reveal competitiveness 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, labor 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, Sreedhar and Ghosh, 2012; Moreno, 2012; Gorre, Magulgad and Ramos, 2012; Belyi and Greene, 2012; Hughes, 2012). As of now, this interpretation is not in the mainstream. Figure 5 shows that the national rent that is generated from the extraction of the natural resource (revenue less cost, as percentage of GDP) negatively relates to the level of government disclosure of data on the economic production in oil, gas and mineral industries.

Figure 5: Public data on natural resource extraction: *Natural resource rent vs. government data disclosure (year=2010; n=40).*



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).

3.7 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 for what and how much to measure diverge. Hardy (2012c, p. 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".

4. 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 well-known digital divide (Hilbert, 2011b) also perpetuates the era of Big Data.

4.1 Infrastructure access

4.1.1 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.

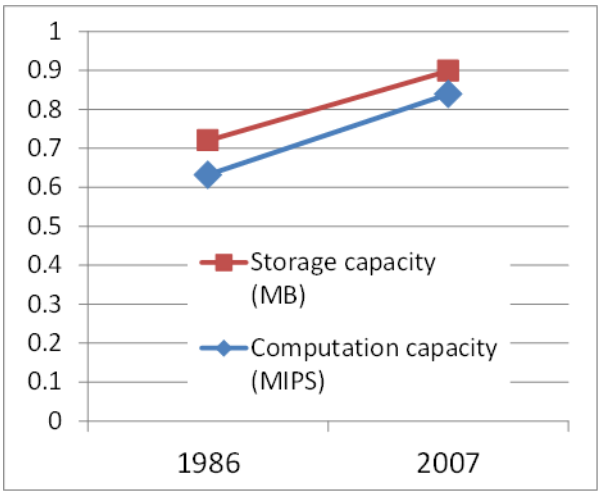
(i) Big Data is still based on samples. While it is the ambition of Big Data to dispense with the need of random sampling techniques by collecting 'everything there is', students of the digital divide are very aware that the online world is only a subsample of everything there is. "Twitter does not represent '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; p. 669). The intensity of the bias is dictated by the intensity of the digital divide. Blumenstock and Eagle (2012) 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" (p.2). Figure 6a 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 from Latin America" with a mobile phone penetration of around 60-80% at the time of the study. Figure 6b shows that the Big Data sample matches official census data impressively well.

Figure 6: Representativeness of Big Data: *comparison of mobile phone subscribers and population at large. (a) Rwanda 2005/09, with mobile phone penetration of 2-20%; (b) 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).

(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, the 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 7 shows this increasing concentration of technological capacity among an ever smaller number of ever more powerful devices in form of the Gini (1921) measure. Naturally, the vast majority of this Big Data hardware capacity resides in highly developed countries.

Figure 7: *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 (Gini = 1 means total concentration with all capacity at one single device; Gini = 0 means total uniformity, with equally powerful devices).*



Source: own elaboration, based on Hilbert (2014a).

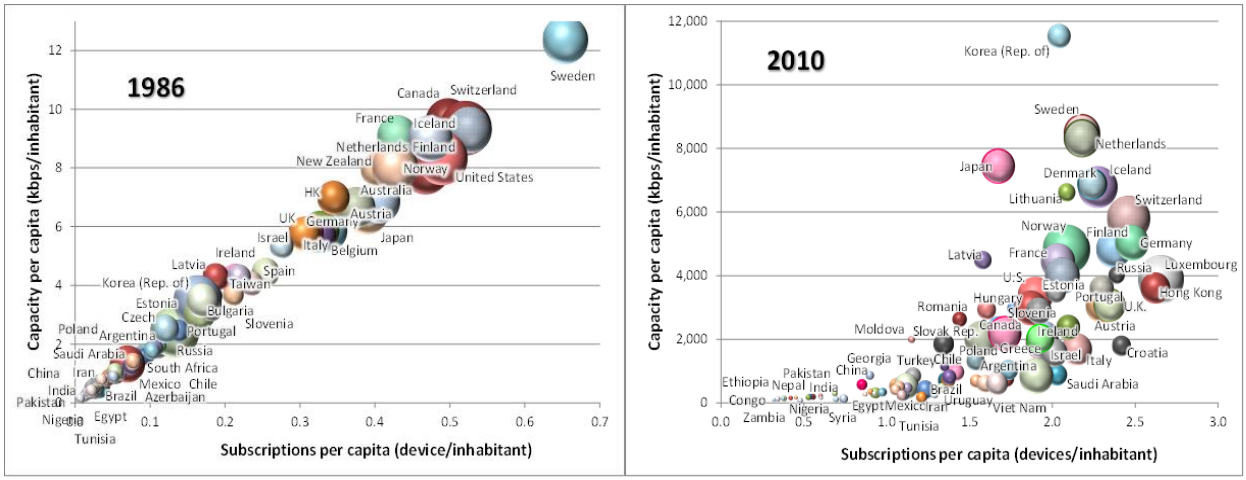
The fundamental condition to convert this increasingly concentrated information capacity among storage and computational devices ("the cloud") into an equalitarian information capacity among and within societies lies in the social ownership of telecommunication access. Telecommunication networks provide a potential technological gateway to the Big Data cloud. Figure 8 shows that this basic condition is ever less fulfilled. Over the past two decades, telecom access has ever become more diversified. In the analog age of 1986, the vast majority of telecom subscriptions were fixed-line phones, and all of them had the same performance. This resulted in a quite linear relation between the number of subscriptions and the average traffic capacity (see Figure 8). Twenty five years later, there's a myriad of different telecom subscriptions with the most diverse range of performances. 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).

4.1.2 Options

One way for developing countries to confront this challenge is to create local hardware capacity by exploiting the decentralized and modular approach inherent to many Big Data solutions. Hadoop, for example, is 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) PC and server hard disks and analyze them in 64 MB blocks, which allows it to "grow with demand while remaining economical at every size" (Shvachko et a;., 2010). With respect to cost-effective distributed computational 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 alternative quite price competitive (López and Hilbert, 2012).

Figure 8: *Subscriptions per capita vs. Capacity per capita (in optimally compressed kbps of installed capacity) for 1986 and 2010. Size of the bubbles represents Gross National Income (GNI) per capita (N = 100).*



Source: own elaboration, based on Hilbert (2014b).

4.2 Generic services

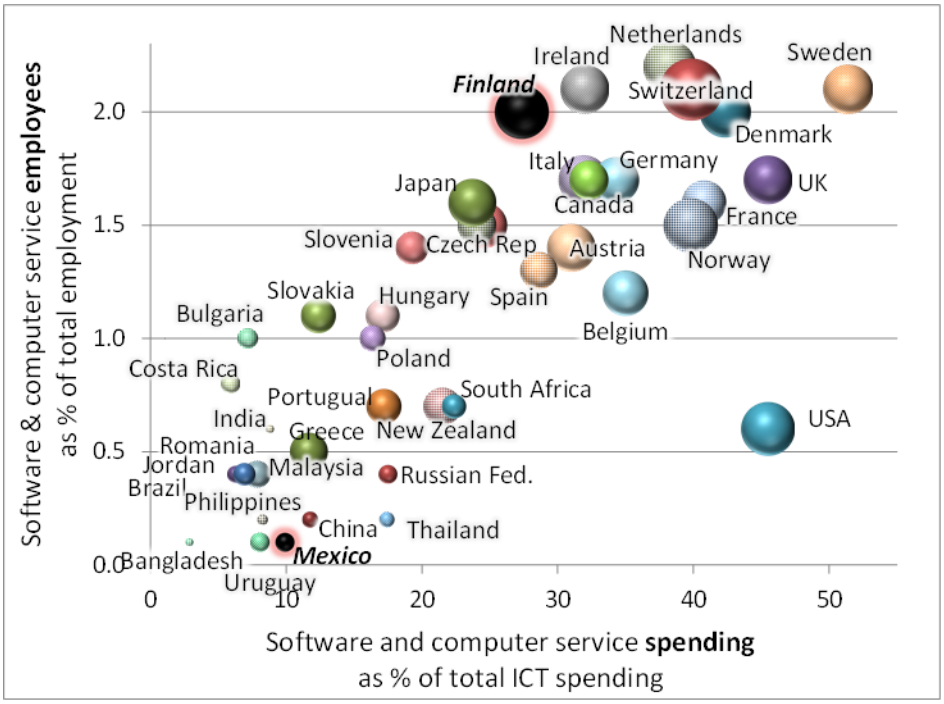
Additional to the tangible hardware infrastructure, Big Data relies heavily on software services to analyze the data. This includes both financial and human resources.

4.2.1 Challenges

Figure 9 shows the shares of software and computer service spending of total ICT spending (horizontal x-axis) and of software and computer service employees of total employees (vertical y-axis) 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 capabilities for software and computer services in relative terms. Envisioning Big Data capabilities in every enterprise, organization 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 specialized in software and computer services (see Finland vs. Mexico in Figure 9), especially when trying to adopt and fine-tune technological solutions to domestic requirements in developing countries (Romijn and Caniëls, 2011).

Figure 9: Spending (horizontal x-axis) and employees (vertical y-axis) of software and computer services (as % of respective total). *Size of bubbles represents total ICT spending per capita (n=42 countries).*



Source: own elaboration, based on UNCTAD, 2012.

4.2.2 Options

There are two basic options on how to obtain Big Data services: in-house or outsourcing. Many organizations opt for a hybrid solution and use on-demand cloud resources to supplement in-house Big Data deployments (Dumbill, 2012), as in-house solutions are alone are notoriously costly (examples coming from large firms like Tesco, Target, Amazon or Wal-Mart). Outsourcing solutions benefit from the extremely high fix-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 specialized data service provider (among the largest being Acxiom, Experian, Epsilon, and InfoUSA). They offer historic voting behavior of politicians, evaluations of customer comments on social ratings sites like Yelp, 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 borderline

dance with the limits of the law and moral practice, they came under the scrutiny of policy makers (U.S. 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 then emptied several savings accounts (Duhigg, 2007). Such obvious crimes are only the pique of an increasing ice-berg of potential discrimination due to Big Data transparency (ranging from personal credit ratings to school acceptance).

Given the commoditization of data, data also becomes subject to existing economic divides. With a global revenue of an estimated US$ 5 -10 billion in 2012/2013 (Feinleib, 2012), the Big Data market has already become bigger than the size of half of the world's national economies in its first years. 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; p. 673-674). 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, as discussed in the chapter of Policy and Strategy.

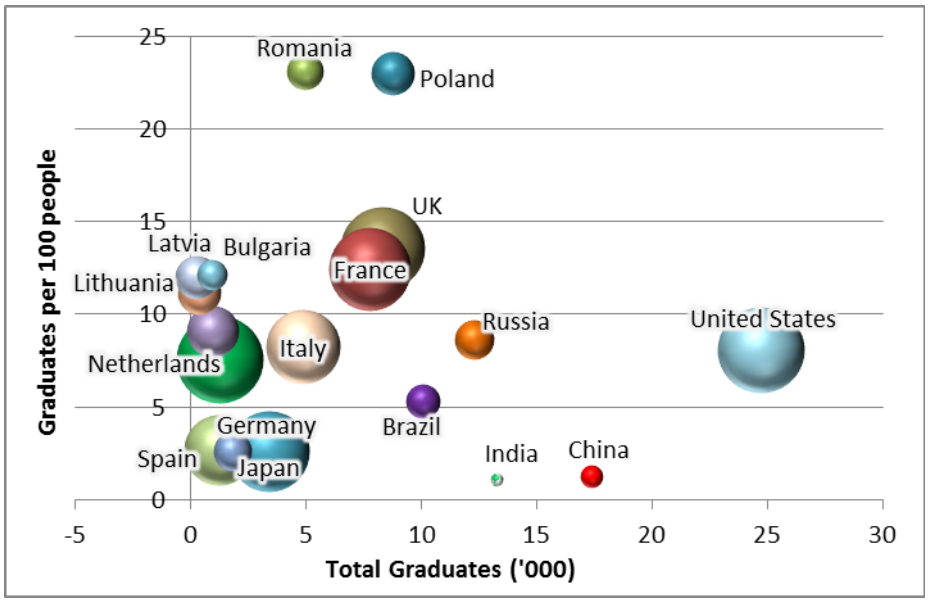
4.3 Capacities and skills

4.3.1 Challenges

Case studies on the use of Big Data applications in development project show that adequate training for data specialists and managers is one of the main reasons for failure (Noormohammad, et al., 2010). Manyika, et al. (2011) predict that by 2018, even the job magnet United States will face a shortage of some 160,000 professionals with deep analytical skills (of a total of 450,000 in demand), as well as a shortage of 1.5 million data managers that are able to make informed decisions based on analytic findings (of a total of 4 million in demand). This shows that there is a global shortage, with a disproportional negative effect on developing countries and the public sector disproportionally (Borne, 2013; Freedman Consulting, 2013). Figure 10 shows that perspectives in this regard are actually mixed for different parts of the developing world. Some countries with relatively low income levels achieve extremely high graduation rates

for professionals with deep analytical skills (high up on the vertical y-axis in Figure 8). In general, countries from the former Soviet bloc (e.g. Romania, Poland, Latvia, Lithuania, and Bulgaria) produce a high level of analysts. The world's large developing BRIC countries (Brazil, Russia, India and China) produce 40 % of the global professionals with deep analytical skills, twice and many as the university power-hose of the United States (far to the right on the x-axis in Figure 10). Traditional leaders of the global economy, such as Germany and Japan, are comparatively ill-equipped to satisfy domestic demand with internal sources. This leads to the long-standing and persistent discussion about brain drain and possible brain circulation in a global diaspora of data-savvy managers, analysts, statisticians, and computer scientists.

Figure 10: Graduates with deep analytical training*: total (horizontal x-axis), per 100 people (vertical y-axis), Gross National Income (GNI) (size of bubbles).*



Source: own elaboration, based on Manyika, et al., 2011 and World Bank, 2010. Note: Counts people taking graduate or final-year undergraduate courses in statistics or machine learning (a subspecialty of computer science).

It is not only the quantity, but also the quality of analytical skills that matters. An inventory of 52 social media studies from 2005-2012 revealed that more than one third of the exercises that claimed to prove the predictive power of social media data did not even run any explicit predictive analytics (but 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.

4.3.2 Options

Most policy implications in this challenge are similar to the well-studied implications of science and technology education and the notorious brain-drain / brain-gain from previous technological revolutions (e.g. Saxenian, 2007). One innovative way of dealing with the shortage of skilled professionals are 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 one quarter of scientists have the necessary skills to analyze available data, while one third said they could obtain the skills through collaboration (Science Staff, 2011). Collaborative setups include Wikis to collectively decode genes or analyze molecular structures (Waldrop, 2008), and aid in the classification of galaxies GalaxyZoo (galaxyzoo.org) and complex protein-folding problems (folding.stanford.edu). The alternative to collaboration is competition. During 2010-2011 the platform Kaggle attracted over 23,000 data scientists worldwide in data analysis competitions with cash prizes between US$ 150 and US$ 3,000,000 (Carpenter, 2011). While the Netflix Prize of US$ 1,000,000 is the much-cited epitome of such analytics competitions (Bell et al., 2010), a more common example includes 57 teams (among others from Chile, Antigua and Barbuda, and Serbia) who helped an Australian to predict the amount of money spend by tourists in a specific area (a valuable insight for a mere US$ 500 cash price) (Hyndman, 2010).

5. Policy and Strategy

No technology, including Big Data, is inherently good or bad for development (Kranzberg, 1986). The maximization of opportunities and the minimization 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 a certain awareness about the importance of data analytics and the urgency to undertake required adjustments. The magnitude of the challenge becomes clear when considering that Ghana's statistical authorities took 17 years to adopt the UN system of national accounts from 1993. In 2010 the 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 reverse effects of well-intended efforts. Emblematic is the case of the digitization of twenty million land records in Bangalore, which created a big data source aimed at benefiting 7 million small farmers from over 27,000 villages (Chawla and Bhatnagar, 2004). Contrary to expectations, the usual dominant players were in a much better position to exploit the provided data, resulting in a perpetuation of existing inequalities (Benjamin et al., 2007).

5.1 Incentives: positive feedback

One kind of intervention consists in positively encouraging and fostering desired outcomes. Two of the most common ways consists in providing funds for data and in 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 U.S. include the earmarked US$ 19 billion of the American Recovery and Reinvestment Act to encourage U.S. physicians to adopt electronic medical recordkeeping systems (Bollier, 2010), and the US$ 700-800 million subsidies of the Office of Cyberinfrastructure (OCI) of the U.S. National Science Foundation (NSF), to be partially invested into "large-scale data repositories and digitized scientific data management systems" (NSF, 2012). Part of the ambition to bring Big Data to the general public consists in investments into data visualization (Frankel and Reid, 2008) (see Figure 11 for a simple example). NSF and the Journal Science have invested into data visualization competitions for over ten consecutive years (Norman, 2012; Science Staff, 2014).

Figure 11: Word cloud of this article: *one simple and quick way to visualize Big Data.*

Source: The full text of this paper; world cloud created using www.Wordle.net ; i.e. http://www.wordle.net/show/wrdl/8112509/BigData4D_article

In developing countries, small financial incentives can already make a large difference. Eagle (2013) reports that a decentralized short-message-system (SMS) to collect blood bank inventory data from Kenyan hospitals initially failed because it did not consider the 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, summing up to a total cost of USD 240 to maintain this life-saving service for 24 Kenyan hospitals.

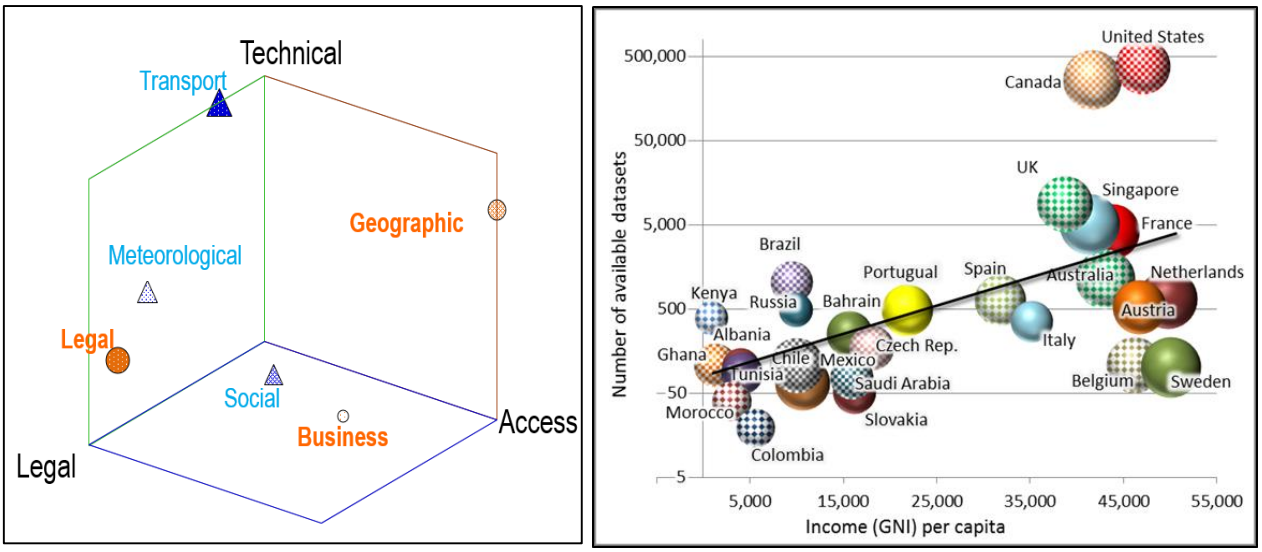
(ii) Open data: Another incentive for Big Data consists in providing open data to the public. This approach sees data as a public good. For example, geospatial data[7] (Vincey, 2012) and weather and climate data (GFDRR, 2012) are among the most widely published public data. The ongoing

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

discussion about the openness of digital data moves along different dimensions (Figure 12a). It includes the use of technical standards of the provided data (e.g. impractical PDF files vs. structured Excel spreadsheets vs. 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 12a). 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 governments" (Lathrop and Ruma, 2010; Kum, Ahalt and Carsey, 2011; Concha and Naser, 2012; WEF and Vital Wave, 2012). It is estimated that each organization of the U.S. government is estimated to host some 1.3 Petabytes of data, compared with a national organizational mean of 0.7 PB, while the government itself hosts around 12 % of the nationally stored data, and the public sector related sectors of education, health care and transportation another 13 % (Manyika, et al., 2011). In other words, if data from the public sector would be openly available, around one quarter of existent 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, Chui, and Manyika, 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 to provide hundreds of datasets on demographics, public expenditures, and natural resources for public access. Also international organization, 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 correlated strongly with the level of economic well-being and perceived transparency of a country (Figure 12b). Several low income countries are more active than their developed counterparts in making databases publicly available (see e.g. Kenya, Russia and Brazil), while other countries with traditionally high perceived transparency are more hesitant (e.g. Chile, Belgium, Sweden).

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

Figure 12: Open data: *(a) Classification of type of public sector information information in Spain along three open data dimensions; (b) Number of datasets provided on central government portal (vertical y-axis, logarithmic scale), Gross National Income per capita (horizontal x-axis),*



Corruption Perception Index (size of bubbles: larger, more transparent) (year=2011; n=27).

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)

4.2 Regulation: negative feedback

The other option to guide the Big Data paradigm into the desired development direction consists in the creation 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 U.S. before the 1920 (Beniger, 1986) and inspired novelists like Huxley (1932) and Orwell (1948).[9] Big Data has taken this issue to a new level. To say it in the words of the editor of the influential computer science Journal "Communications of the ACM": "NSA [U.S. National Security Agency], indeed, is certainly making Orwell's surveillance

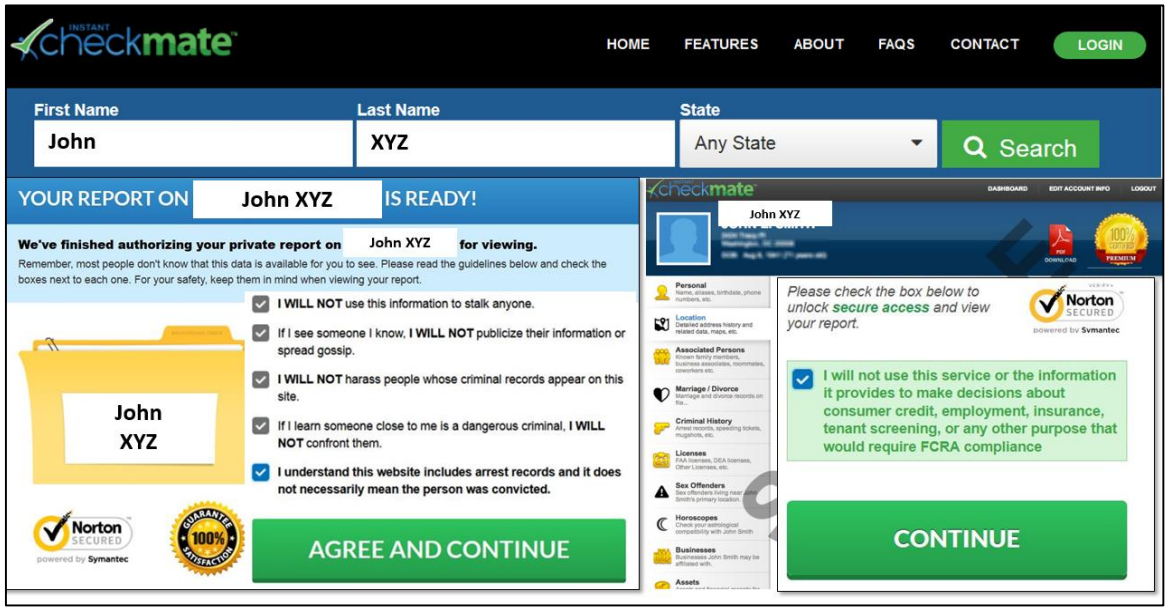
[9] "By comparison with that existing today, all the tyrannies of the past were half-hearted and inefficient" (Orwell, 1948; 2, 9).

technology seem rather primitive in comparison" (Vardi, 2013; p.5). Fact of the matter is that digital information always leaves a potential trace that can be tracked and analyzed (Andrews, 2012) and that "any data on human subjects inevitably raise privacy issues" (Nature Editorial, 2007; p. 637). One common distinction is 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, or tweets, etc.

With regard to regulation of one or the other kind of data, a 2014 White House report suggest to basically give up on the futile attempt to regulate which data may and 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; p.xiii). Those adverse consequences would then be severely punished (such as in the above-cited example of discrimination and criminal acts against elderly (see Duhigg, 2007)). This is in line the legal philosophy of not regulating the possession of something, but their illegal usage. For example, the case of fire arms in the United States follows this legal philosophy. Besides foregoing practical challenges in the regulation of possession, the advantage of permitting possession consists in assuring not to curb any of the potential benefits of its usage. This well-known discussion about benefits, dangers and practicality is currently being held with regard to Big Data.

The currently existing legal grey zone is exemplified by the fact that several Big Data provider 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. Figure 13 shows that before obtaining the purchased report, the consumer has to click to consent 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" (Figure 13). It seems difficult to control for such commitments without more binding legal rules or control. At the same time, the self-reported data abuses of the NSA (U.S. Government, 2012) demonstrated that even when rules and regulations are in place (incomplete as they might be), the benefits of breaking those rules are so tempting that governments have admittedly undertaken illegal steps to take advantage of it. These issues are much less regulated in developing countries, be it in academia, the private sector or government agencies.

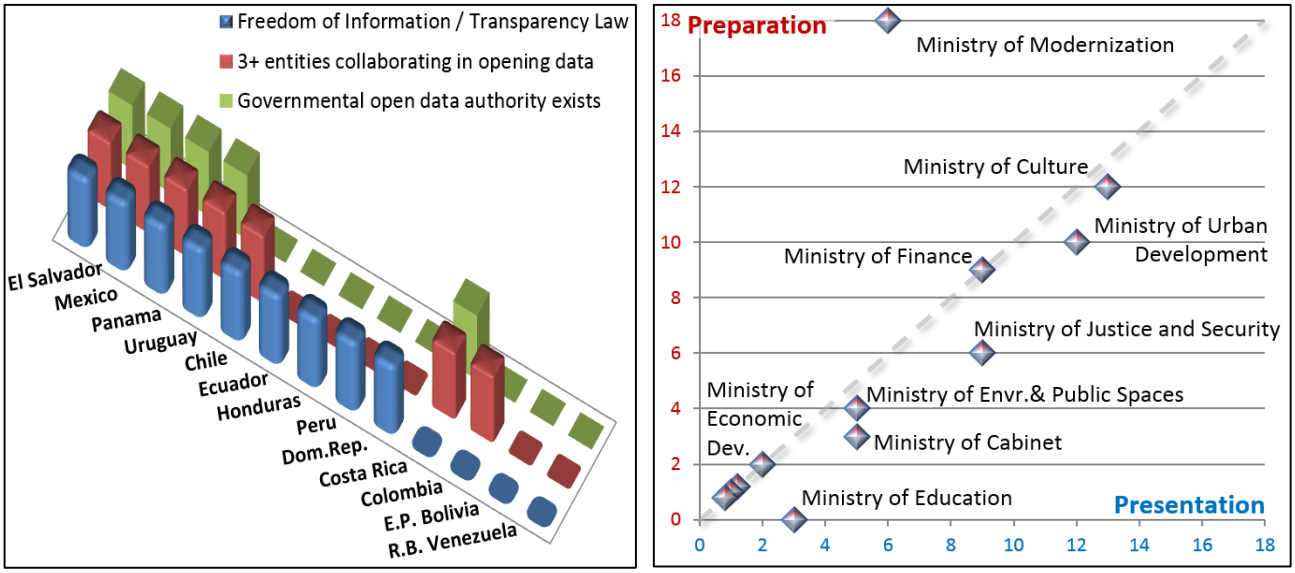
Figure 13: Checkmate: *screenshot of consumer consent to obey privacy norms.*



Source: http://www.instantcheckmate.com .

(ii) Regulating openness. Open government data does not need to rely on voluntary and incentivizing 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 catalyzer for the provision of open government data. The devil in this matter is in the implementation. Figure 14a shows that at least four countries in Latin America count with such legislation, but not with any active authority or collaboration that pushes the implementation of open data forward. On the contrary, Costa Rica and Colombia do not count with such legislation, but with concrete institutional projects. Figure 14b illustrates that the pinpointed assignation of a leading authority can accelerate the implementation of open government directives. The specialized Ministry of Modernization of the city government of Buenos Aires has prepared and published the lion's share of the available databases (18 of 67), while it then provides them to other authorities which present them on their webpages.

Figure 14: Open data theory and praxis: *(a) Existence of open data legislation and implementation in Latin America; (b) Number of datasets prepared and presented by government authorities in Buenos Aires.*



Source: own elaboration, based on (a) Jadue (2012); (b) Belbis (2012).

(iii) Interoperability. While one of the main opportunities of Big Data is data fusion (see above), bringing data from different sources together is also one of its main challenges. Large parts of valuable data lurk in data silos of different departments, regional offices, and specialized 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 standardized. The regulation of data interoperability standards has become a pressing issue for the Big Data paradigm in both developed (NSF, 2007), as well as in developing countries (UN-ECLAC, 2007; de la Fuente, 2012).

6. 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 evaluators have been overtaken by machines in many fields. Only a decade ago we would have still be surprised if somebody would have told us that artificial intelligence diagnostics outperform human aneurysms radiologists with a success rate of 95% versus 70 % (Raihan, 2010). We got used to it

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

quickly. When fostering this kind of approach, we inevitably give a lot of power to algorithms (Baker, 2008). Per definition, algorithms can only execute processes that are programmed into them on the basis of data that is given to them. Unfortunately, both the programming of algorithms and the nature of the data have short-comings.

First, the programmer rarely is 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 US$ 23 million on Amazon.com by competing algorithms that forecasted 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 70-75 % of trading volume in the U.S. in 2009 (Hendershott, Jones and Menkveld, 2011) and has triggered several unreasonable sell-offs at stock markets (triggering a so-called "flash-crash") (Kirilenko, et al., 2011, Steiner, 2012).

Second, all statistics are informed by data, which is inevitably from the past (in the best case, the 'real-time past', which 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 the purpose to create a future that is different from the past. This limits the insights obtained from Big Data (Hilbert, 2014d; 2014e). 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 at providing 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 disgracefully fail 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 to explore 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 does not only change empirical data

science, but also theory-driven modelling that allows to explore scenarios that never existed, for example through computer simulation. The most powerful candidate are so-called agent-based models (e.g. 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 limitations as all empirical work: it is exclusively post-factum.

7. Conclusion

In a 2014 White House report of 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 emphasizing 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 reconfirmed that the Big Data paradigm holds both promises and perils 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 health care, security, economic productivity, and disaster- and resource management, among others. The extraction of actionable knowledge from the vast amounts 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 innovations is never immediate and uniform, but inescapably creates divides while the diffusion process lasts. This review has shown that the Big Data paradigm currently runs through such unequal diffusion process that is compromised by structural characteristics, such as the lacks of infrastructure, human capital, economic resource availability, and institutional frameworks in developing countries. This inevitably 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 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 certainly not a panacea. It is indispensable to accompany and guide this transition with proactive policy options and pin-pointed development projects.

[11] Given the size of the phenomenon, the literature review is non-exhaustive. It has been realized during the period of 2012 – 2014, mainly by the identification of academic articles and reports through the specialized online search engine Google Scholar.

References

Abella, A. (2014). *Modelling the Economic impact of information reuse in Spain* (Master Thesis in Organización de Empresas). Universidad Rey Juan Carlos. http://calendariolibre.com/gobernamos/wp-content/uploads/2013/11/Modelling_economic_impact_information_reuse_in_Spain.pdf

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

Aguilar Sánchez, C. (2012). *Brazil: No Easy Miracle. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Transparency and Accountability Initiative. http://www.revenuewatch.org/sites/default/files/Brazil_TAI.pdf

Althouse, B. M., Ng, Y. Y., & 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, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, (Science: Discoveries). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Andrews, L. (2012). I Know Who You Are and I Saw What You Did: Social Networks and the Death of Privacy. Simon and Schuster.

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

Baker, S. (2008). *The Numerati* (First Edition.). Houghton Mifflin Harcourt.

Banko, M., & 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* (pp. 26–33). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1073012.1073017

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., & Volinsky, C. (2010). All Together Now: A Perspective on the Netflix Prize. *CHANCE*, *23*(1), 24–29. doi:10.1080/09332480.2010.10739787

Belyi, A., & Greene, S. (2012). *Russia: A Complex Transition. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Transparency and Accountability Initiative. http://www.revenuewatch.org/sites/default/files/Russia_TAI_eng.pdf

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & 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*. Harvard University Press.

Benjamin, S., Bhuvaneswari, R., & Manjunatha, P. R. (2007). *Bhoomi: "E-Governance," or, An anti-politics machine necessary to globalize Bangalore*. A CASUM-m Working Paper. http://casumm.files.wordpress.com/2008/09/bhoomi-e-governance.pdf

Berners-Lee, T. (2006, July 27). Linked Data. http://www.w3.org/DesignIssues/LinkedData.html

Blumenstock, J., Eagle, N., & Fafchamps, M. (2012). Risk and Reciprocity Over the Mobile Phone Network: Evidence from Rwanda. http://archive.nyu.edu/handle/2451/31441

Blumenstock, J. E., & Eagle, N. (2012). Divided We Call: Disparities in Access and Use of Mobile Phones in Rwanda. *Information Technologies & International Development*, *8*(2), pp. 1–16.

Blumenstock, J. E., Gillick, D., & Eagle, N. (2010). Who's Calling? Demographics of Mobile Phone Use in Rwanda. In *AAAI Spring Symposium: Artificial Intelligence for Development*. AAAI. http://dblp.uni-trier.de/db/conf/aaaiss/aaaiss2010-1.html#BlumenstockGE10

Bollier, D. (2010). *The promise and peril of Big data*. Washington D.C.: The Aspin Institute. http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(90003), 7280–7287. doi:10.1073/pnas.082080899

Borbora, Z., Srivastava, J., Kuo-Wei Hsu, & Williams, D. (2011). Churn Prediction in MMORPGs Using Player Motivation Theories and an Ensemble Approach. In *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on Social Computing (SocialCom)* (pp. 157–164). IEEE. doi:10.1109/PASSAT/SocialCom.2011.122

Borne, K. (2013, June 11). Big Data, Small World. Presented at the TEDxGeorgeMasonU. http://www.youtube.com/watch?v=Zr02fMBfuRA&feature=youtube_gdata_player

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of "big data"? *McKinsey Quarterly*, *McKinsey Global Institute*(October). https://www.mckinseyquarterly.com/Are_you_ready_for_the_era_of_big_data_2864

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN eLibrary*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486

Buckee, C. O., Wesolowski, A., Eagle, N., Hansen, E., & Snow, R. W. (2013). Mobile phones and malaria: modeling human and parasite travel. *Travel Medicine and Infectious Disease*, *11*(1), 15–22. doi:10.1016/j.tmaid.2012.12.003

Butler, D. (2007). Data sharing threatens privacy. *Nature News*, *449*(7163), 644–645. doi:10.1038/449644a

Carpenter, J. (2011). May the Best Analyst Win. *Science*, *331*(6018), 698–699. doi:10.1126/science.331.6018.698

Castells, M. (2009). *The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume I* (2nd ed.). 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* (pp. 91–115). Oxford: Westview Press.

Chawla, R., & 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. http://info.worldbank.org/etools/docs/reducingpoverty/case/96/fullcase/India%20Bhoomi%20Full%20Case.pdf

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165–1188.

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, (275), 314–347.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*, 2–9. doi:10.1111/j.1475-4932.2012.00809.x

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

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, *86*(1), 39–45. doi:10.4269/ajtmh.2012.11-0597

Concha, G., & 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. http://www.eclac.org/ddpe/publicaciones/xml/9/46119/W465.pdf

Darthmouth. (2012). *Unwarranted Variations and Their Remedies: Findings from the Dartmouth Atlas of Health Care*. Lectures. http://www.dartmouthatlas.org/pages/multimedia

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. http://www.eclac.org/ddpe/publicaciones/xml/9/46119/W465.pdf

De Mauro, A., Greco, M., & 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. doi:10.13140/2.1.2341.5048

Devarajan, S. (2011). Africa's statistical tragedy. http://blogs.worldbank.org/africacan/africa-s-statistical-tragedy

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

Driscoll, K. (2012). From Punched Cards to "Big Data": A Social History of Database Populism. *Communication 1*, *1*(1). http://scholarworks.umass.edu/cpo/vol1/iss1/4

Duhigg, C. (2007, May 20). Bilking the Elderly, With a Corporate Assist. *The New York Times*. http://www.nytimes.com/2007/05/20/business/20tele.html

Dumbill, E. (2012, January 11). What is big data? An introduction to the big data landscape. *O'Reilly Radar*. http://radar.oreilly.com/2012/01/what-is-big-data.html

Dutta, R., Sreedhar, R., & Ghosh, S. (2012). *India: Development at a Price. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Transparency and Accountability Initiative. http://www.revenuewatch.org/sites/default/files/India_TAI_eng.pdf

Eagle, N. (2013). People are not cookies. TEDxEast. https://www.youtube.com/watch?v=AT2q17EhGBM&feature=youtube_gdata_player

Epstein, J. M., & Axtell, R. L. (1996). *Growing Artificial Societies: Social Science from the Bottom Up* (First.). A Bradford Book.

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of the ACM (Association for Computing Machinery)*, *48*(11), 87–92. doi:10.1145/1096000.1096010

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, *460*(7256), 685–686. doi:10.1038/460685a

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

FOI (Freedom of Information). (2012). Freedom of information legislation - Wikipedia, the free encyclopedia. Retrieved March 28, 2012, from http://en.wikipedia.org/wiki/Freedom_of_information_legislation

Frankel, F., & Reid, R. (2008). Big data: Distilling meaning from data. *Nature*, *455*(7209), 30. doi:10.1038/455030a

Freedman Consulting. (2013). *A Future or Failure?: The Flow of Technology Talent into Government and Civil Society*. Ford Foundation and McArthur Foundation. www.fordfoundation.org/pdfs/news/afutureoffailure.pdf

Freeman, C., & Louçã, F. (2002). *As Time Goes By: From the Industrial Revolutions to the Information Revolution*. Oxford University Press, USA.

Frias-Martinez, V., Frias-Martinez, E., & Oliver, N. (2010). A Gender-centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records. In *AAAI 2010 Spring Symposia Artifical Intelligence for Development*.

Frias-Martinez, V., & Virseda, J. (2013). Cell Phone Analytics: Scaling Human Behavior Studies into the Millions. *Information Technologies & International Development*, *9*(2), pp. 35–50.

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

Gell-Mann, M., & 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). https://www.gfdrr.org/opendri

Gilbert, N., & Troitzsch, K. (2005). *Simulation for the Social Scientist* (2 edition.). Maidenhead, England ; New York, NY: Open University Press.

Gini, C. (1921). Measurement of Inequality of Incomes. *The Economic Journal*, *31*(121), 124–126.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. doi:10.1038/nature07634

Goffman, E. (1959). *The Presentation of Self in Everyday Life* (1st ed.). Anchor.

Goldenberg, A., Shmueli, G., Caruana, R. A., & 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), 5237–5240. doi:10.1073/pnas.042117499

Gorre, I., Magulgad, E., & Ramos, C. A. (2012). *Phillippines: Seizing Opportunities. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Transparency and Accountability Initiative. http://www.revenuewatch.org/sites/default/files/Philippines_TAI.pdf

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The Predictive Power of Online Chatter. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 78–87). New York, NY, USA: ACM. doi:10.1145/1081870.1081883

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

Hardy, Q. (2012a). Bizarre Insights From Big Data. http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/

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

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

Helbing, D., & Balietti, S. (2010). From Social Data Mining to Forecasting Socio-Economic Crisis. *arXiv:1012.0178*. http://arxiv.org/abs/1012.0178

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does Algorithmic Trading Improve Liquidity? *The Journal of Finance*, *66*(1), 1–33. doi:10.1111/j.1540-6261.2010.01624.x

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), 756–770. doi:10.1016/j.worlddev.2009.11.019

Hilbert, M. (2011a). Video animation on the word's technological capacity to store, communicate and compute information. June 7th-8th 2011, Santa Clara California: The Economist, proof-inc. http://www.youtube.com/watch?v=iIKPjOuwqHov

Hilbert, M. (2011b). The end justifies the definition: The manifold outlooks on the digital divide and their practical usefulness for policy-making. *Telecommunications Policy*, *35*(8), 715–736. doi:10.1016/j.telpol.2011.06.012

Hilbert, M. (2011c). Mapping the dimensions and characteristics of the world's technological communication capacity during the period of digitization. In Working Paper. Mauritius: International Telecommunication Union (ITU). http://www.itu.int/ITU-D/ict/wtim11/documents/inf/015INF-E.pdf

Hilbert, M. (2012). Towards a Conceptual Framework for ICT for Development: Lessons Learned from the Latin American "Cube Framework." *Information Technologies & International Development*, *8*(4, Winter; Special issue: ICT4D in Latin America), 243–259 (Spanish version: 261–280).

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), 856–861. doi:10.1002/asi.23031

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), 821–835. doi:10.1002/asi.23020

Hilbert, M. (2014d). The ultimate limitation of big data for development. *SciDev*, (Opinion). http://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, x=independently organized TED talks.

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, *332*(6025), 60 –65. doi:10.1126/science.1200970

Hilbert, M., & López, P. (2012). How to Measure the World's Technological Capacity to Communicate, Store and Compute Information? Part I: results and scope. *International Journal of Communication*, *6*, 956–979.

Hubbard, D. W. (2011). *Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities* (1st ed.). Wiley.

Hughes, T. (2012). *South Africa: A Driver of Change. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Transparency and Accountability Initiative. http://www.revenuewatch.org/sites/default/files/SouthAfrica_TAI.pdf

Huxley, A. (1932). *Brave new world* (1st edition.). London: Chatto & Windus.

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

IBM. (2011). *Vestas: Turning climate into capital with big data* (Case study). http://public.dhe.ibm.com/common/ssi/ecm/en/imc14702usen/IMC14702USEN.PDF

IBM News. (2007, August 16). Beacon Institute and IBM Team to Pioneer River Observatory Network [News release]. Retrieved March 7, 2012, from http://www-03.ibm.com/press/us/en/pressrelease/22162.wss

IBM News. (2009a, May 13). IBM Ushers In Era Of Stream Computing [News release]. Retrieved March 7, 2012, from http://www-03.ibm.com/press/us/en/pressrelease/27508.wss

IBM News. (2009b, November 19). UMBC Researchers Use IBM Technology to Fight Rising Threats of Forest Fires [News release]. Retrieved March 7, 2012, from http://www-03.ibm.com/press/us/en/pressrelease/28863.wss#release

Isafiade, O., & 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* (pp. 21:1–21:2). New York, NY, USA: ACM. doi:10.1145/2537052.2537073

ITU (International Telecommunication Union). (2012). *Measuring the Information Society 2012*. Geneva: International Telecommunication Union, ITU-D. http://www.itu.int/ITU-D/ict/publications/idi/

Jadue, V. (2012). *Estudio sobre Datos Abiertos Gubernamentales en America Latina y el Caribe*. Chile: United Nations ECLAC, IDRC, OD4D.

James, J. (2012). *Data never sleeps: How much data is generated every minute?*. http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559. doi:10.1108/IntR-06-2012-0114

Kelly, K. (2011, March 28). *Keynote Web 2.0 Expo SF 2011*. San Francisco. http://www.web2expo.com/webexsf2011/public/schedule/detail/19292

King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, *331*(6018), 719–721. doi:10.1126/science.1197872

Kirilenko, A. A., Kyle, A. S., Samadi, M., & Tuzun, T. (2011). The Flash Crash: The Impact of High Frequency Trading on an Electronic Market. *SSRN Electronic Journal*. doi:10.2139/ssrn.1686004

Kranzberg, M. (1986). Technology and History: "Kranzberg's Laws." *Technology and Culture*, *27*(3), 544. doi:10.2307/3105385

Kum, H.-C., Ahalt, S., & Carsey, T. M. (2011). Dealing with Data: Governments Records. *Science*, *332*(6035), 1263–1263. doi:10.1126/science.332.6035.1263-a

Lathrop, D., & Ruma, L. (2010). *Open Government: Collaboration, Transparency, and Participation in Practice* (1st ed.). O'Reilly Media.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. (2010). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *Winter 2011*. http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. doi:10.1126/science.1248506

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … Alstyne, M. V. (2009). Computational Social Science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Letouzé, E. (2012). *Big Data for Development: Opportunities and Challenges* (White p). New York: United Nations Global Pulse. http://www.unglobalpulse.org/projects/BigDataforDevelopment

Lohr, S. (2009, August 6). For Today's Graduate, Just One Word: Statistics. *The New York Times*. http://www.nytimes.com/2009/08/06/technology/06stats.html

López, P., & Hilbert, M. (2012). *Methodological and Statistical Background on The World's Technological Capacity to Store, Communicate, and Compute Information* (online document). http://www.martinhilbert.net/WorldInfoCapacity.html

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, *109*(29), 11576–11581. doi:10.1073/pnas.1203882109

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

Manovich, L. (2012). Trending: The Promises and the Challenges of Big Social Data. In M. Gold (Ed.), *Debates in the Digital Humanities* (pp. 460–476). Minneapolis: The University of Minnesota Press. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

MarketPsych. (2014). Thomson Reuters MarketPsych Indices (TRMI). Retrieved September 6, 2014, from https://www.marketpsych.com/data/

Martínez, E. F., & Martínez, V. F. (2014, January 30). Method, computer programs and a use for the prediction of the socioeconomic level of a region. http://www.google.com/patents/US20140032448

Masuda, Y. (1980). *The Information Society as Post-Industrial Society*. Tokyo: World Future Society.

Michener, R. G. (2009). *The Surrender of Secrecy? Explaining the Strength of Transparency and Access to Information Laws* (SSRN Scholarly Paper No. ID 1449170). Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=1449170

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE*, *8*(4), e61981. doi:10.1371/journal.pone.0061981

Moreno, R. (2012). *Mexico: A Moment of Opportunity. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Transparency and Accountability Initiative. http://www.revenuewatch.org/sites/default/files/Mexico_TAI_eng.pdf

Moumni, B., Frias-Martinez, V., & 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* (pp. 1199–1208). New York, NY, USA: ACM. doi:10.1145/2494091.2497350

Naef, E., Muelbert, P., Raza, S., Frederick, R., Kendall, J., & Gupta, N. (2014). *Using Mobile Data for Development*. Cartesian and Bill & Melinda Gates Foundation. https://docs.gatesfoundation.org/Documents/Using%20Mobile%20Data%20for%20Development.pdf

Nature Editorial. (2007). A matter of trust. *Nature*, *449*(7163), 637–638. doi:10.1038/449637b

Nature Editorial. (2008). Community cleverness required. *Nature*, *455*(7209), 1–1. doi:10.1038/455001a

Nelson, S. (2008). Big data: The Harvard computers. *Nature*, *455*(7209), 36. doi:10.1038/455036a

Noormohammad, S. F., Mamlin, B. W., Biondich, P. G., McKown, B., Kimaiyo, S. N., & 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. doi:10.1016/j.ijmedinf.2010.01.002

Norman, C. (2012). 2011 International Science & Engineering Visualization Challenge. *Science*, *335*(6068), 525–525. doi:10.1126/science.335.6068.525

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

NSF (National Science Foundation). (2012b). Community-based Data Interoperability Networks (INTEROP). Retrieved March 28, 2012, from http://www.nsf.gov/od/oci/about.jsp

Open Government Partnership. (2014). Open Government Declaration. http://www.opengovpartnership.org/sites/www.opengovpartnership.org/files/page_files/OGP_Declaration.pdf

O'Reilly Radar. (2011). *Big Data Now: Current Perspectives from O'Reilly Radar*. OReilly Media - A.

Orwell, G. (1948). *1984*. The Literature Network, Jalic LLC,. http://www.online literature.com/orwell/1984

Overeem, A., Leijnse, H., & Uijlenhoet, R. (2013). Country-wide rainfall maps from cellular communication networks. *Proceedings of the National Academy of Sciences*, *110*(8), 2741–2745. doi:10.1073/pnas.1217961110

Paul, C. K., & Mascarenhas, A. C. (1981). Remote Sensing in Development. *Science*, *214*(4517), 139–145. doi:10.1126/science.214.4517.139

Paul, M., & 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* (pp. 265–272). Association for the Advancement of Artificial Intelligence. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2880/3264

Peres, W., & Hilbert, M. (2010). *Information Societies in Latin America and the Caribbean Development of Technologies and Technologies for Development*. Santiago: United Nations ECLAC. http://www.cepal.org/publicaciones/xml/3/43803/Libro_Cepal_98.pdf

Perez, C. (2004). Technological Revolutions, Paradigm Shifts and Socio-Institutional Change. In E. Reinert (Ed.), *Globalization, Economic Development and Inequality: An alternative Perspective* (pp. 217–242). Cheltenham: Edward Elgar. http://www.carlotaperez.org/papers/basic-technologicalrevolutionsparadigm.htm

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

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An Emerging Tool for Social Scientists. *Sociological Methods & Research*, *37*(3), 426–454. doi:10.1177/0049124108330005

Raihan, I. (2010). Managing Big data. http://www-03.ibm.com/systems/resources/systems_Managing_Big_Data_Podcast_Transcription.pdf

Revenue Watch Institute, & Transparency International. (2010). *Revenue Watch Index 2010. Transparency: Governments and the oil, gas and mining industries*. Transparency International. http://www.revenuewatch.org/rwindex2010/pdf/RevenueWatchIndex_2010.pdf

Rissanen, J. (2010). *Information and Complexity in Statistical Modeling* (Softcover reprint of hardcover 1st ed. 2007.). Springer.

Ritterfeld, U., Shen, C., Wang, H., Nocera, L., & Wong, W. L. (2009). Multimodality and Interactivity: Connecting Properties of Serious Games with Educational Outcomes. *CyberPsychology & Behavior*, *12*(6), 691–697. doi:10.1089/cpb.2009.0099

Ritterman, J., Osborne, M., & Klein, E. (2009). Using prediction markets and Twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Media*.

Romijn, H. A., & Caniëls, M. C. J. (2011). Pathways of Technological Change in Developing Countries: Review and New Agenda. *Development Policy Review*, *29*(3), 359–380. doi:10.1111/j.1467-7679.2011.00537.x

Saxenian, A. (2007). *The New Argonauts: Regional Advantage in a Global Economy*. Harvard University Press.

Schumpeter, J. (1939). *Business Cycles: A Theoretical, Historical, And Statistical Analysis of the Capitalist Process*. New York: McGraw-Hill. http://classiques.uqac.ca/classiques/Schumpeter_joseph/business_cycles/schumpeter_business_cycles.pdf

Science Staff. (2011). Challenges and Opportunities. *Science*, *331*(6018), 692–693. doi:10.1126/science.331.6018.692

Science Staff. (2014). 2013 Visualization Challenge. *Science*, *343*(6171), 600–610. doi:10.1126/science.343.6171.600

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Shapiro, C., & Varian, H. R. (1998). *Information Rules: A Strategic Guide to the Network Economy* (1st ed.). Harvard Business Press.

Shen, C., & Williams, D. (2011). Unpacking Time Online: Connecting Internet and Massively Multiplayer Online Game Use With Psychosocial Well-Being. *Communication Research*, *38*(1), 123–149. doi:10.1177/0093650210377196

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. doi:10.1214/10-STS330

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Q.*, *35*(3), 553–572.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)* (pp. 1–10). doi:10.1109/MSST.2010.5496972

Simon, H. A. (2002). Science seeks parsimony, not simplicity: searching for pattern in phenomena. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, Inference and Modelling: keeping it sophisticatedly simple* (pp. 32–72). Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511493164

Slavin, K. (2011). *How algorithms shape our world* (Vol. Talks). http://www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world.html

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of Predictability in Human Mobility. *Science*, *327*(5968), 1018–1021. doi:10.1126/science.1177170

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of Socioeconomic Levels Using Cell Phone Records. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *User Modeling, Adaption and Personalization* (pp. 377–388). Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-22362-4_35

Statista. (2014). Statistics and Market Data on Mobile Internet & Apps. Retrieved September 3, 2014, from www.statista.com/markets/424/topic/538/mobile-internet-apps/

Steiner, C. (2012). Automate This: how algorithms came to rule our world. Gildan Media LLC.

Tan-Mullins, M. (2012). *China: Gradual Change. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Transparency and Accountability Initiative. http://www.revenuewatch.org/sites/default/files/China_TAI_eng.pdf

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

Toole, J. L., Eagle, N., & Plotkin, J. B. (2011). Spatiotemporal Correlations in Criminal Offense Records. *ACM Trans. Intell. Syst. Technol.*, *2*(4), 38:1–38:18. doi:10.1145/1989734.1989742

Transparency International. (2011). *Corruption Percpetion Index 2011*. http://www.transparency.org/cpi2011

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. doi:10.1126/science.7455683

UNCTAD (United Nations Conference on Trade and Development). (2012). *Information Economy Report 2012: The Software Industry and Developing Countries*. Geneva: UNCTAD. http://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=271

UNDP (United Nations Development Programme). (2014). Human Development Index (HDI). Retrieved September 6, 2014, from http://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. http://www.eclac.org/ddpe/publicaciones/xml/7/37347/WhiteBook.pdf

U.S. Government. (2012). Memorandum of the Chief of SID Oversight & Compliance (OC-034-12) (Memorandum) (p. 13). http://apps.washingtonpost.com/g/page/national/nsa-report-on-privacy-violations-in-the-first-quarter-of-2012/395/

U.S. Senate. (2013). *A Review of the Data Broker Industry: Collection, Use , and Sale of Consumer Data for Marketing Purposes*. Committee on Commerce, Science, and Transportation. http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577

Vardi, M. Y. (2013). The end of the American network. *Communications of the ACM*, *56*(11), 5–5. doi:10.1145/2524713.2524714

Vincey, C. (2012, July). *Opendata benchmark: FR vs UK vs US*. Presented at the Dataconnexions launch conference, Google France. http://www.slideshare.net/cvincey/opendata-benchmark-fr-vs-uk-vs-us

Waldrop, M. (2008). Big data: Wikiomics. *Nature News*, *455*(7209), 22. doi:10.1038/455022a

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. In S. J. Yang, A. M. Greenberg, & M. Endsley (Eds.), *Social Computing, Behavioral - Cultural Modeling and Prediction* (pp. 231–238). Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-29047-3_28

WEF (World Economic Forum), & Vital Wave Consulting. (2012). Big Data, Big Impact: New Possibilities for International Development. Retrieved August 24, 2012, from http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development

Wennberg, J. E. (2011). Time to tackle unwarranted variations in practice. *BMJ*, *342*(mar17 3), d1513–d1513. doi:10.1136/bmj.d1513

Wennberg, J. E., O'Connor, A. M., Collins, E. D., & 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), 1564–1574. doi:10.1377/hlthaff.26.6.1564

Wesolowski, A., Stresman, G., Eagle, N., Stevenson, J., Owaga, C., Marube, E., … Buckee, C. O. (2014). Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. *Scientific Reports*, *4*. doi:10.1038/srep05678

White House. (2014a). *Big Data and Privacy: a technological perspective*. Washington D.C.: Executive Office of the President, President's Council of Advisors on Science and Technolo. http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

White House. (2014b). *Big Data: Seizing Opportunities, preserving values*. Exceutive Office of the President. http://www.whitehouse.gov/issues/technology/big-data-review

Zikopoulos, P., Eaton, C., deRoos, D., Detusch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (IBM.). New York: McGraw. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CPM=is_bdebook1&cmp=109HF&S_TACT=109HF38W&s_cmp=Google-Search-SWG-IMGeneral-EB-0508