Big Data for Development:
From Information- to Knowledge Societies

Martin Hilbert (Dr. PhD.)
United Nations Economic Commission for Latin America and the Caribbean (UN ECLAC)
Annenberg School of Communication, University of Southern California (USC)
Email: martinhilbert@gmail.com

Abstract

The article uses an established three-dimensional conceptual framework to systematically review literature and empirical evidence related to the prerequisites, opportunities, and threats of Big Data Analysis for international development. On the one hand, the advent of Big Data delivers the cost-effective prospect to improve decision-making in critical development areas such as health care, employment, economic productivity, crime and security, and natural disaster and resource management. This provides a wealth of opportunities for developing countries. On the other hand, all the well-known caveats of the Big Data debate, such as privacy concerns, interoperability challenges, and the almighty power of imperfect algorithms, are aggravated in developing countries by long-standing development challenges like lacking technological infrastructure and economic and human resource scarcity. This has the potential to result in a new kind of digital divide: a divide in data-based knowledge to inform intelligent decision-making. This shows that the exploration of data-based knowledge to improve development is not automatic and requires tailor-made policy choices that help to foster this emerging paradigm.

Acknowledgements: The author thanks Canada's International Development Research Centre, Canada (IDRC) for commissioning a more extensive study that laid the groundwork for the present article. He is also indebted with Manuel Castells, Nathan Petrovay, Francois Bar, and Peter Monge for food for thought, as well as Matthew Smith, Rohan Samarajiva, Sriganesh Lokanathan, and Fernando Perini for helpful comments on draft versions.

Table of Contents

The ability to "cope with the uncertainty caused by the fast paced 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 purely 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 probabilistic estimates, our uncertainty will be reduced on average. This is not merely a narrative analogy, but a well-established proven mathematical theorem of information theory that provides the foundation for all kinds of statistical and probabilistic analysis (Cover and Thomas, 2006; p. 29; also Rissanen, 2010).[3]

The Big Data[4] paradigm (Nature Editorial, 2008) provides loads of additional data to fine-tune the models and estimates that inform all sorts of decisions. This amount of additional information stems from unprecedented increases in (a) information flow, (b) information storage, and (c) information processing.

(a) During the two decades of digitization, the world's effective capacity to exchange information through two-way telecommunication networks grew from 0.3 exabytes in 1986 (20 % digitized) to 65 exabytes two decades later in 2007 (99.9 % digitized) (Hilbert and López, 2011). In contrary to analog information, digital information inherently leaves a trace that can be analyzed (in real-time or later on). In an average minute of 2012, Google receives around 2,000,000 search queries, Facebook users share almost 700,000 pieces of content, and Twitter users send roughly 100,000 microblogs (James, 2012). Additional to these mainly human-generated telecommunication flows, surveillance cameras, health sensors, and the "Internet of things" (including household appliances and cars) are adding a large chunk to ever increasing data streams (Manyika, et al., 2011).

[1] Reality is as complex that we never know all conditions and processes and always need to abstract from it in models on which to base our decisions. Everything excluded from our limited model is seen as uncertain "noise". Therefore: "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).

[3] In information-theoretic terms we would say that every probability is a conditional probability (conditioned on some initial distribution; Caves, 1990) and that conditioning (on more realizations of the conditioning variable) reduces entropy (uncertainty) on average: $H(X|Y) \geq H(X|YZ)$ (see Cover and Thomas, 2006; p. 29). Note that we have to condition on real information (not "miss-information") and that this theorem holds on average (it might be that one particular piece of information increases uncertainty, such as specific evidence in court, etc.).

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

(b) At the same time, our technological memory roughly doubled every 40 months (about every three years), growing from 2.5 optimally compressed exabytes in 1986 (1 % digitized), to around 300 optimally compressed exabytes in 2007 (94 % digitized) (Hilbert and López, 2011; 2012). In 2010, it costs merely US$ 600 to buy a hard disk that can store all the world's music (Kelly, 2011). This increased memory has to capacity to ever store a larger part of an incessantly growing information flow. In 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) 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).[5] What fuels this expectation is the fact that our capacity to compute information in order to make sense of data has grown two to three times as fast as our capacity to store and communicate information over recent decades: while our storage and telecommunication capacity has grown at some 25-30% per year over recent decades, our capacity to compute information has grown at some 60-80% annually (Hilbert and López, 2011, 2012). Our computational capacity has grown from 730 tera-IPS (instructions per seconds) in 1986, to 196 exa-IPS in 2007 (or roughly 2*10^20 instructions per second; which is roughly 500 times larger since the number of seconds since the big bang) (Hilbert and López, 2012).

As such, the crux of the "Big Data" paradigm is actually not the increasingly large amount of data itself, but its analysis for intelligent decision-making (in this sense, the term "Big Data Analysis" would actually be more fitting than the term "Big Data" by itself). 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, 2011) to "Knowledge Societies":

[5] In the Big Data world, a distinction is often made between structured data, such as the traditional kind that is produced by questionnaires, or "cleaned" by artificial or human supervisors, and unstructured raw data, such the data produced by online and Web communications, video recordings, or sensors.

building on the digital infrastructure that led to vast increases in information, the current challenge consists in converting this digital information into knowledge that informs intelligent decisions.

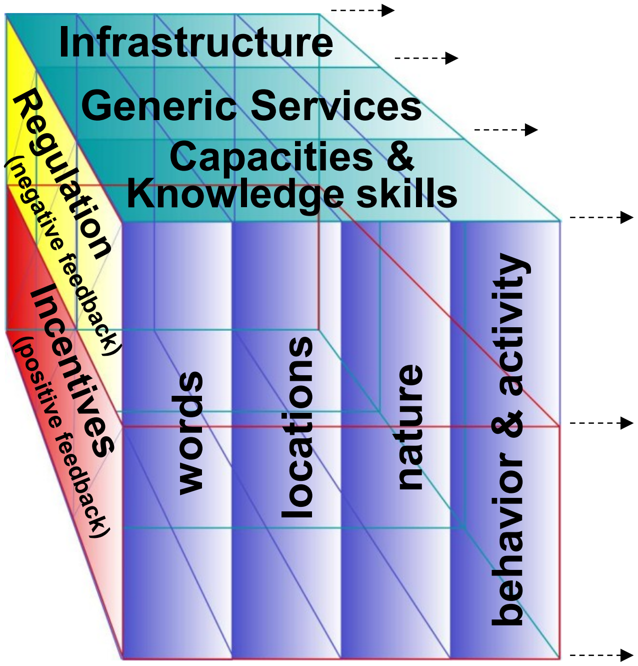
The extraction of knowledge from databases is not new by itself. Driscoll (2012) distinguishes between three historical periods: early mass-scale computing (e.g. the 1890 punched card based U.S. Census that processed some 15 million individual records), the massification of small personal databases on microcomputers (replacing standard office filing cabinets in small business during the 1980s), and, more recently, the emergence of both highly centralized systems (such as Google, Facebook and Amazon) and the interconnection of uncountable small databases. The combination of sufficient bandwidth to interconnect decentralized data producing entities (be they sensors or people) and the computational capacity to process the resulting storage provides huge potentials for improving the countless smaller and larger decisions involved in any development dynamic. In this article we systematically review existing literature and related empirical evidence to obtain a better understanding of the opportunities and challenges involved in making the Big Data Analysis paradigm work for development.

Conceptual Framework

In order to organize the available literature and empirical evidence, we use an established three-dimensional conceptual framework that models the process of digitization as an interplay between technology, social change, and guiding 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 Analysis.

The first requisites of making Big Data work for development are a solid technological (hardware) infrastructure, generic (software) services, and human capacities and skills. These horizontal layers are used to analyze different aspects and kinds of data, such as words, locations, nature's elements, and human behavior, among others. While this set-up is necessary for Big Data Analysis, it is not sufficient for development. In the context of this article, (under)development is broadly understood as (the deprivation of) capabilities (Sen, 2000). Rejecting pure technological determinism, all technologies (including ICT) are normatively neutral and can also be used to 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 help us create better public and private goods and services, rather than leading to increased State and corporate control that poses a threat to societies (especially those with fragile and incipient institutions)? Not 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 in 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). The result is a three-dimensional framework, whereas different circumstances (e.g. infrastructure deployment) and strategies (e.g. regulations) intersect and affect different aspects of Big Data Analysis.

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



In this article we will work through the different aspects of this framework. We will start with some examples of Big Data for development through the tracking of words, locations, nature's elements, and human behavior and economic activity. After this introduction to the ends of Big Data, we will look at the means, specifically the current distribution of the current hardware infrastructure and software services among developed and developing countries. We will also spend a considerable amount of time of the distribution of human capital and will go deeper into the specific skill requirements for Big Data. Last but not least, we will review aspects and examples of regulatory and incentive systems for the Big Data paradigm.

Applications of Big Data for Development

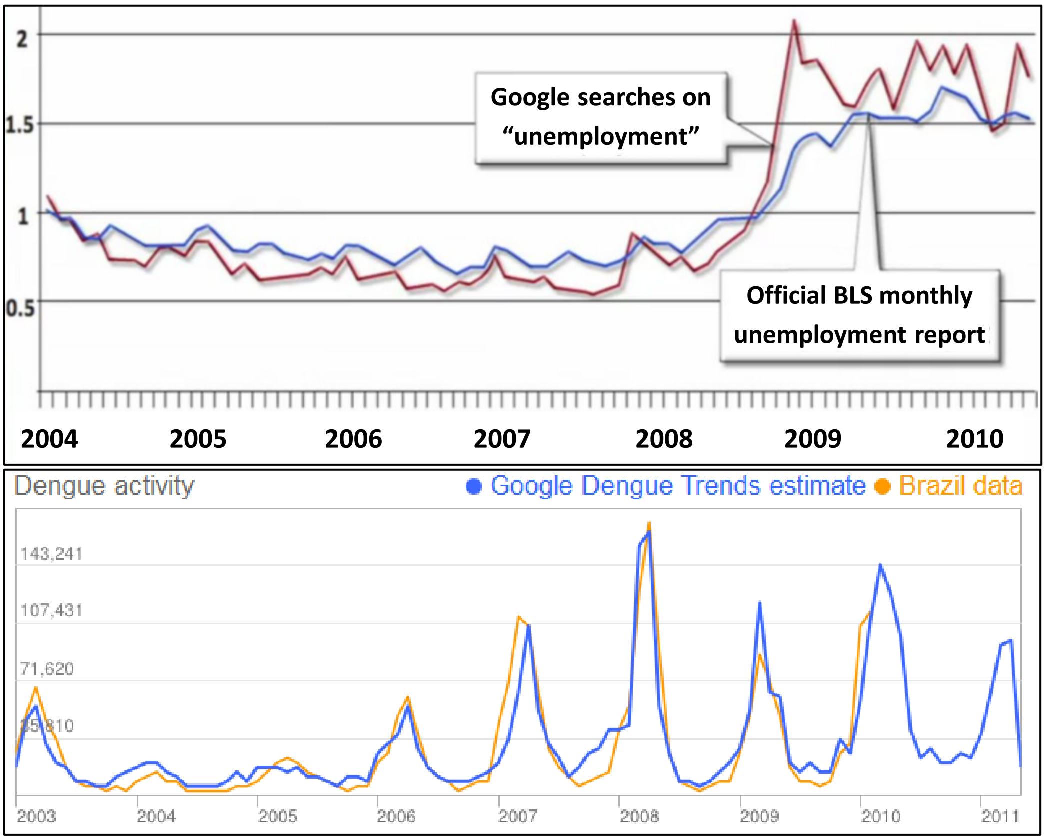
From a macro-perspective, it is expected that Big Data informed decision-making will have a similar positive effect on efficiency and productivity as ICT have had during the recent decade (see Brynjolfsson and Hitt, 1995; Jorgenson, 2002; Melville, Kraemer, and Gurbaxani, 2004; Castells, 2009; Peres and Hilbert, 2010). However, it is expected to add to the existing effects of digitization. Brynjolfsson, Hitt, and Kim (2011) surveyed 111 large firms in the U.S. in 2008 about the existence and usage of data for business decision making and for the creation of a new products or services. They found that firms that adopted Big Data Analysis have output and productivity that is 5 – 6 % higher than what would be expected given their other investments and information technology usage. Measuring the storage capacity of organizational units of different sectors in the U.S. economy, the consultant company McKinsey (Manyika, et al., 2011) shows that this potential goes beyond data intensive banking, securities, investment and manufacturing sectors. Several sectors with particular importance for development are quite data intensive: education, health, government, and communication host one third of the data in the country. The following reviews some illustrative case studies in development relevant fields like employment, crime, water supply, and health and disease prevention.

Tracking words

One of the most readily available and most structured kinds of data 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). Analyzing comments, searches or online posts can produce nearly the same results for statistical inference as household surveys and polls. 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). Using a similar logic, Google was able to spot trends in the Swine Flu epidemic in January 2008 roughly two weeks before the U.S. Center of Disease Control (O'Reilly Radar, 2011). Given this amount of free data, the work- and time-intensive need for statistical sampling seems almost obsolete.

The potential for development is straightforward. Figure 2b illustrates the match between the data provided publicly by the Ministry of Health about dengue and the corresponding Google Trend data, which is able to make predictions were official data is still lacking. In another application, an analysis of the 140 character long microblogging service Twitter showed that it contained important information about the spread of the 2010 Haitian cholera outbreak and was up available up to two weeks earlier than official statistics (Chunara, Andrews and Brownstein, 2012). The tracking of words can be combined with other databases, such as done by Global Viral Forecasting, which specializes in predicting and preventing pandemics (Wolfe, Gunasekara and Bogue, 2011), or the World Wide Anti-Malarial Resistance Network that collates data to inform and respond rapidly to the malaria parasite's ability to adapt to drug treatments (Guerin, Bates and Sibley, 2009).

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

Tracking locations

Location-based data are usually obtained from four primary sources: in-person credit or debit card payment data; in-door tracking devices, such as RFID tags on shopping carts; GPS chips in mobile devices; or cell-tower triangulation data on mobile devices. The last two provide the largest potential, especially for developing countries, which already own three times more mobile phones than their developed counterparts (reaching a penetration of 85 % in 2011 in developing countries) (ITU, 2011). By 2020, more than 70 percent of mobile phones are expected to have GPS capability, up from 20 percent in 2010 (Manyika, et al., 2011), which means that developing countries will produce the vast majority of location-based data.

Location-based services have obvious applications in private sector marketing, but can also be put to public service. In Stockholm, for example, a fleet of 2,000 GPS-equipped vehicles, consisting of taxis and trucks, provide data in 30 - 60 seconds intervals in order to obtain 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 to historical data, combining it with weather forecasts, and information from past traffic patterns, etc. Such traffic analysis does not only save time and gasoline for citizens and businesses, but is also useful for public transportation, police and fire departments, and, of course, road administrators and urban planners.

Chicago Crime and Crimespotting in Oakland present robust interactive mapping environments that allow users to track instances of crime and police beats in their neighborhood, while examining larger trends with time-elapsed visualizations. Crimespotting pulls daily crime reports from the city's Crimewatch service and tracks larger trends and provide user-customized services such as neighborhood-specific alerts. The system has been exported and successfully implemented in other cities.

Tracking nature

One of the biggest sources of uncertainty is nature. Reducing this uncertainty through data analysis can quickly lead to tangible impacts. A recent project by the United Nations University uses climate and weather data to analyze "where the rain falls" in order to improve food security in developing countries (UNU, 2012). A global beverage company was able cut its beverage inventory levels by about 5 % by analyzing rainfall levels, temperatures, and the number of hours of sunshine (Brown, Chui, and Manyika, 2011, p. 9). Combing Big Data of nature and social practices, 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). Real cost reduction means increasing productivity and therefore economic growth.

The same tools can be used to prevent downsides and mitigate risks that stem from the environment, such as natural disasters and resource bottlenecks. 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 during wildfires (IBM News, Nov. 2009). This allows local fire and safety officials to make more informed decisions on public evacuations and health warnings and provides them with real-time forecasts. Similarly, the Open Data for Resilience Initiative fosters the provision and analysis of data from climate scientists, local governments and communities to reduce the impact of natural disasters by empowering decisions-makers in 25 (mainly developing) countries with better information on where and how to build safer schools, how to insure farmers against drought, and how to protect coastal cities against future climate impacts, among other intelligence (GFDRR, 2012).

Sensors, robotics and computational technology have also been used to track river and estuary ecosystems, which help officials to monitor water quality and supply through the movement of chemical constituents and large volumes of underwater acoustic data that tracks the behavior of fish and marine mammal species (IBM News, May 2009). For example, the River and Estuary Observatory Network (REON) allows for minute-to-minute monitoring of the 315-mile New York's Hudson River, monitoring this important natural infrastructure for 12 million people who depend on it (IBM News, 2007). 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 is reported to have improved emergency response time by 30 % (IBMSocialMedia, 2012).

The optimization of a systems performance and the mitigation of risks are often closely related. The economic viability of alternative and sustainable energy production often hinges on timely information about wind and sunshine patterns, since it is extremely costly to create energy buffers that step in when conditions are not continuously favorable (which they never are). 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).

Tracking behavior

Half a century of game theory has shown that social defectors are among the most disastrous drivers of social inefficiency. The default of trust and the systematic abuse of social conventions are two main behavioral challenges for society. A considerable overhead is traditionally added to social transactions in order to mitigate the risk of defectors. This can be costly and inefficient. Game theory also teaches us that social systems with memory of past and predictive power of future behavior can circumvent such inefficiency (Axelrod, 1984). Big Data can provide such memory and are already used to provide short-term payday loans that are up to 50 %

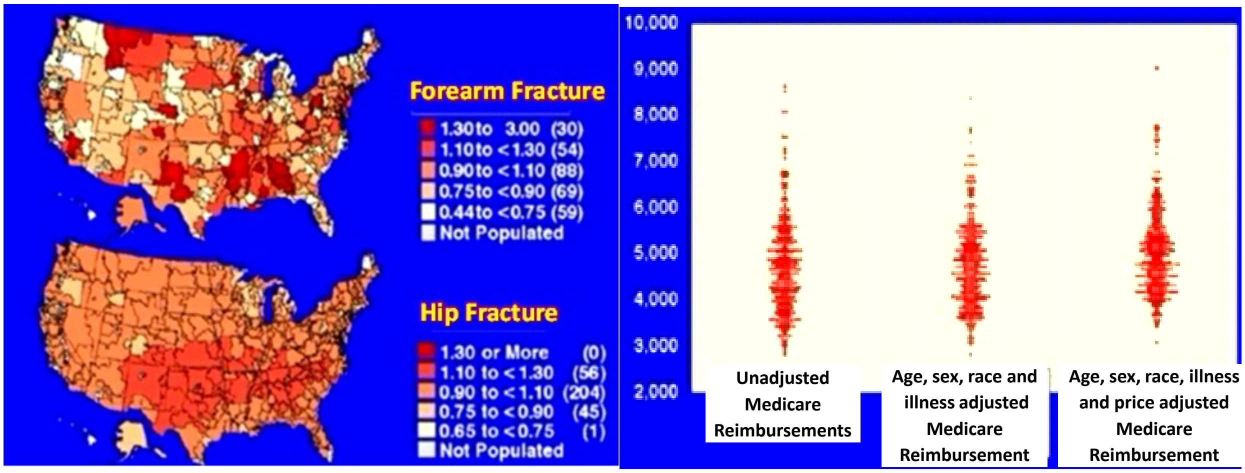
cheaper than the industry's average, judging risk via criteria like cellphone bills and the way how applicants read the loan application Website (Hardy, 2012a).

Behavioral abnormalities are usually spotted by analyzing variations in the behavior of individuals in light of the collective behavior of the crowd. As an example from the health sector, Figure 3a presents the hospitalization rates for forearm- and hip-fractures across the U.S. (Darthmouth, 2012). While the case for hip-fractures is within expected standard deviations (only 0.3 % of the regions show extreme values in the case of hip-fractions), forearm fracture hospitalization rate is 9 times larger (30 % of the regions can be found in the extreme values). The analysis of such variations is often at the heart of Big Data analysis. In this case, four types of variations can generally be found:

- Environmental differences: hip-fractions show a clear geographic pattern in mid-west of the U.S., which could be a reflection of weather, work and alimentation. In practice these variations account for a surprisingly small part of the detected data patterns: Figure 3b shows that the differences in total Medicare spending among regions in the U.S. (which ranges from less than US$ 3000 per patient to almost US$ 9000) is not get reduced when adjusting for demographical differences (age, sex, race), differences in illness patterns, and differences in regional prices.
- Medical errors: some regions systematically neglect preventive measures, and others have an above average rate of mistakes.
- Biased judgment: the need for surgery—one of the main drivers of health care cost—is often unclear, and systematic decision-making biases are common (Wennberg, et al., 2007).
- Overuse and oversupply: Procedures are prescribed simply because the required resources are abundantly available in some regions. The number of prescribed procedures correlates strongly with resource availability, but not at all with health outcomes (Darthmouth, 2012): more health care spending does not reduce mortality (R^2 = 0.01, effectively no correlation); does not affect the rates of elective procedures (R^2 = 0.01), and does not even reduce the level of underuse of preventive measures (R^2 = 0.01); but does lead to a detectable positive correlation with more days in hospital (R^2 = 0.28); with more surgeries during last 6 years of life (R^2 = 0.35); and with visits to medical specialists (R^2 = 0.46) or ten or more physicians (R^2 = 0.43).

With Big Data, a simple analysis of variations allows to detect "unwarranted variations" like the last three, which originate with the underuse, overuse, or misuse of medical care (Wennberg, 2011). These affect the means of health care, but not its ultimate end.

Figure 3: (a) *Patterns of variations in the hospitalization for forearm and hip-fracture across U.S.;* ***(b)*** *Patterns of Medicare Spending U.S.*



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

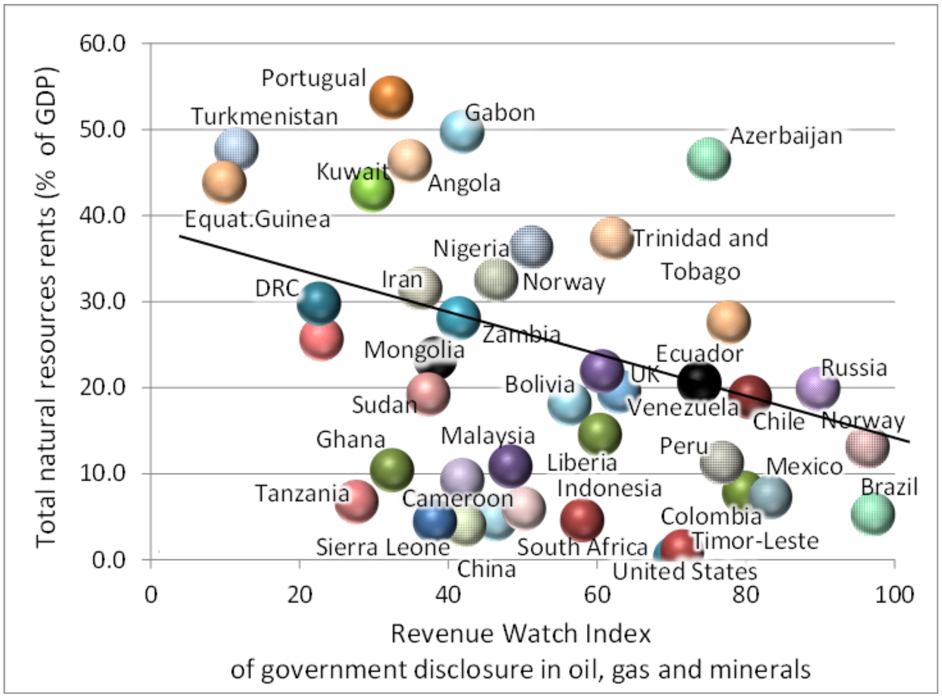
Behavioral data can also be produced by digital applications. Examples of behavioral data generating solutions are online games like World of Warcraft (11 million players in 2011) and FarmVille (65 million users in 2011). Students of multi-player online games can readily predict who is likely to leave the game, explain why that person left, and make suggestions how to provide incentives to keep them playing (Borbora, Srivastava, Hsu and Williams, 2012).

By now, multiplayer online games are also used to track and influence behavior at the same time. Health insurance companies are currently developing multi-layer online games that aim at increasing the fitness levels of their clients. Such games are fed with data from insurance claims and medical records, and combine data from the virtual world and the real world. Points can be earned by checking into the gym or ordering a healthy lunch. The goal is to reduce health care cost, and to increase labor productivity and quality of life (Petrovay, 2012). In order to make this idea work, Big Data solutions recognize that people are guided by dissimilar incentives, such as competing, helping out or leading in a social or professional circle of peers. The collected data allows the incentive structure of the game to adapt to these psychological profiles and individually change peer pressure structures. In order to identify those incentive structures it is essential to collect different kinds of data on personal attributes and behavior, as well as on the network relations among individuals. The tracking of who relates to whom quickly produces vast amounts of data on social network structures, but defines the dynamics of opinion leadership and peer pressure, which are extremely important inputs for behavioral change (e.g. Valente and Saba, 1998).

Tracking economic activity

A contentious area of Big Data for development is the reporting of economic activity that could potentially harm economic competitiveness. 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 of data that relate to the economic activity of these sectors could help to remind the current shortcomings, 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 for now, Figure 4 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 activity in oil, gas and mineral industries: the higher the economic share of resource extraction, the lower the availability of respective data.

Figure 4: 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; 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).

Tracking other data

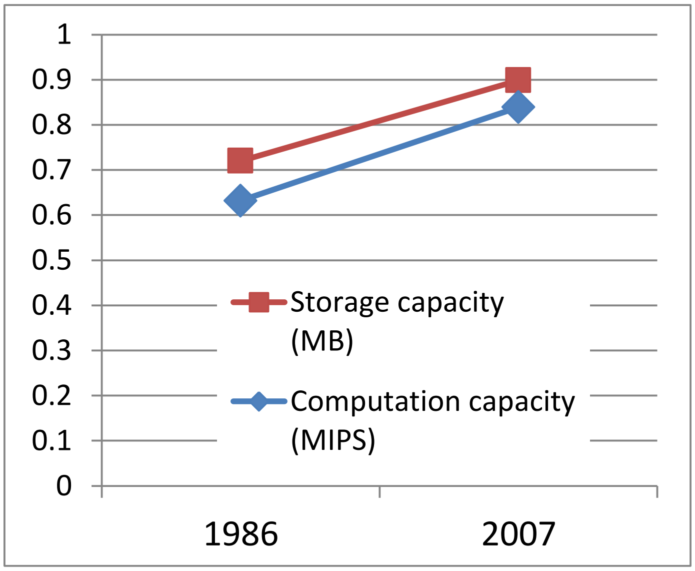
As indicated in the conceptual framework of Figure 1, these are merely illustrative examples of Big Data Analysis. Information is a "difference which makes a difference" (Bateson, 2000; p. 272), and a countless number of variations in data patterns can lead to informative insights. Some additional sources might include the tracking of differences in the supply and use of financial or natural resources, food and aliments, education attendance and grades, waste and exhaust, public and private expenditures and investments, among many others. Current ambitions for what and how much to measure diverge. Hardy (2012b) 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" (p. 4).

What they all have in common is that the longstanding laws of statistics still apply. For example, while large amount of data make the sampling error irrelevant, this does not automatically make the sample representative. For example, boyd and Crawford (2012) underline that "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" (p. 669). We also have to consider that digital conduct is often different from real world conduct. 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). Therefore, studying digital traces might not automatically give us insights into offline dynamics. Besides these biases in the source, the data-cleaning process of unstructured Big Data frequently introduces additional subjectivity.

Infrastructure

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, 2011) also perpetuates the era of Big Data. From a Big Data perspective, it is important to recognize that digitization increasingly concentrated informational storage and computational resources in the so-called "cloud". While in 1986, the top performing 20 % of the world's 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 (see also author, elsewhere). Figure 5 shows the Gini (1921) measure of this increasing concentration of technological capacity among an ever smaller number of ever more powerful devices.

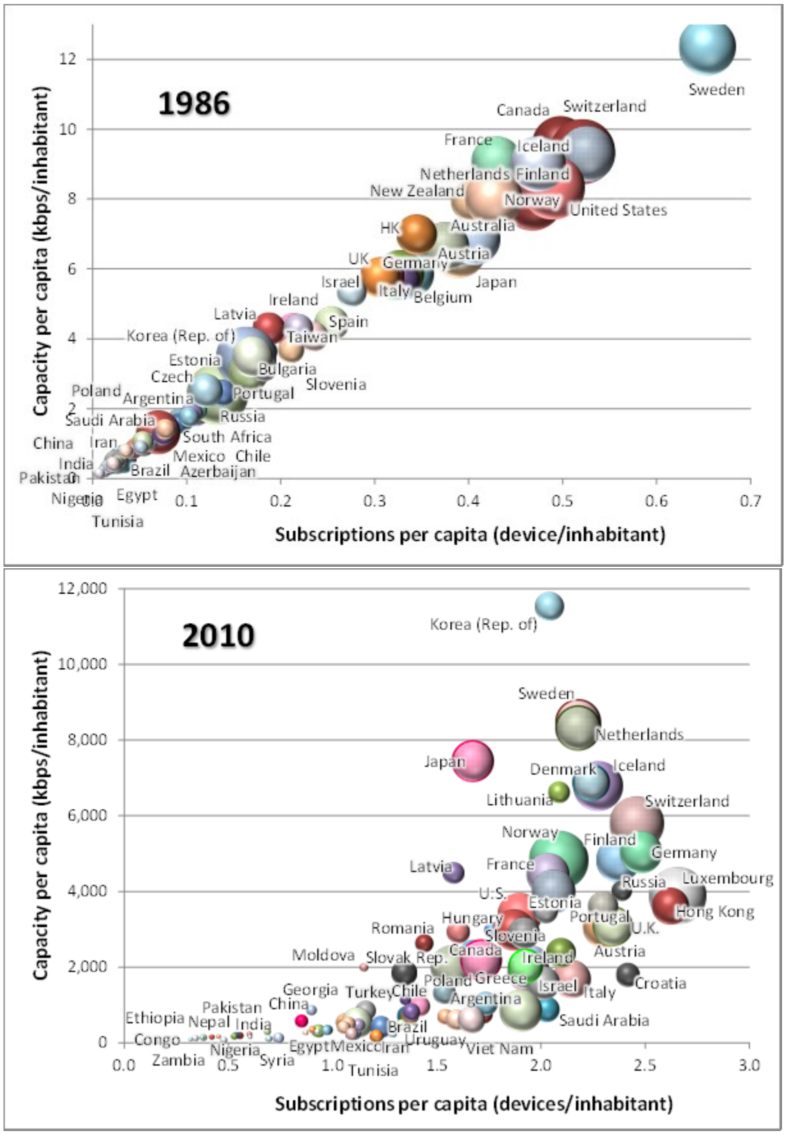
Figure 5: *Gini measure of the world's number of storage and computational devices, and their technological capacity (in optimally compressed MB, and MIPS), 1986 and 2007 (Gini = 1 means total concentration with all capacity at one single device; Gini = 0 means total uniformity, with equally powerful devices).*



Source: own elaboration, for details see author, elsewhere.

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 uniform gateway to the Big Data cloud. Figure 6 shows that this basic condition is ever less fulfilled. Over the past two decades, telecom access has ever become more diversified. 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. 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 6). Twenty years later, there's a myriad of different telecom subscriptions with the most diverse range of performances. This results in a two-dimensional diversity among societies with more or less subscriptions, and with more or less telecommunication capacity.

Figure 6: *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, for details see author, elsewhere.

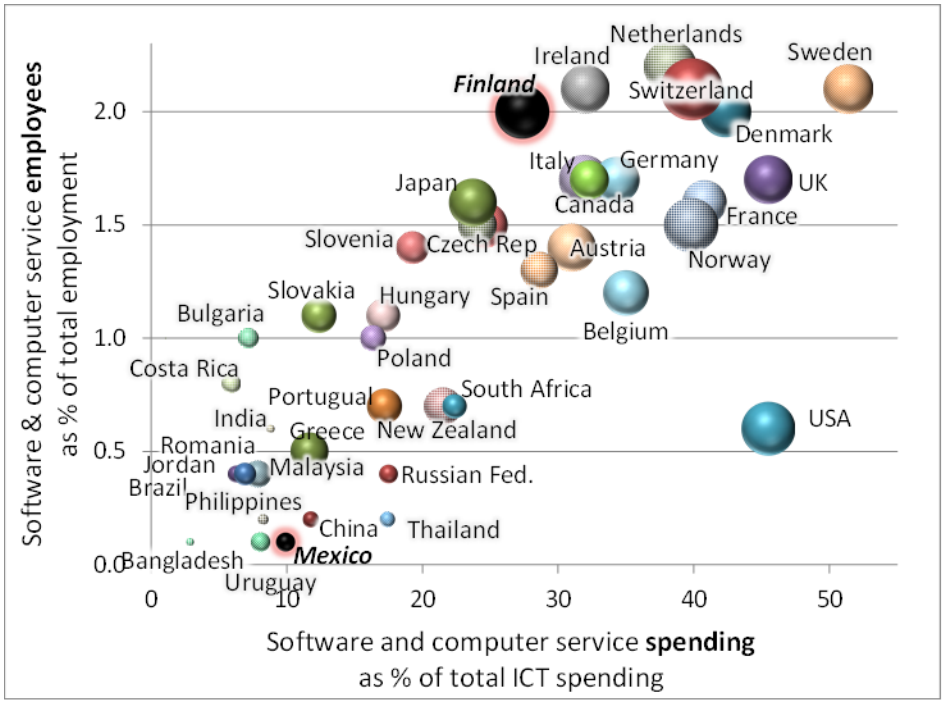
Summing up, incentives inherent to the information economy, such as economies of scale and short product lifecycles (Shapiro and Varian, 1998) increasingly concentrate information storage and computational infrastructure in a “Big Data cloud”. Naturally, the vast majority of this Big Data hardware capacity resides in highly developed countries. The access to these concentrated information and computation resources is skewed by a highly unequal distribution of telecommunication capacities to access those resources. Far from being closed, the digital divide incessantly evolves through an ever changing heterogeneous collection of telecom bandwidth capacities (author, elsewhere). It is important to notice that Figure 6 merely measures the installed telecommunication bandwidth and not actual traffic flows. Considering economic limitations of developing countries, it can be expected that the actual traffic flow is actually more skewed than the installed telecommunication bandwidth.

One way to confront this dilemma consists in creating local Big Data hardware capacity in developing countries. Modular and decentralized approaches seem to be a cost effective alternative. Hadoop, for example, is prominent open-source top-level Apache data-mining warehouse, with a thriving community (the Big Data industry leaders, such as IBM and Oracle embrace Hadoop as an integral part of their products and services). 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. Built in redundancy provide stability even if several of the source drives fail (Zikopoulos, et al., 2012). With respect to computational power, clusters of videogame consoles are frequently used as a substitute for supercomputers for Big Data Analysis (e.g. Gardiner, 2007; Dillow, 2010). Our numbers suggest that 500 PlayStation 3 consoles amount to the average performance of a supercomputer in 2007, which makes this alternative quite price competitive (author, elsewehreSUPP).

Generic Services

Additional to the tangible hardware infrastructure, Big Data relies heavily on software services to analyze the data. Basic capabilities in the production, adoption and adaptation of software products and services are a key ingredient for a thriving Big Data environment. This includes both financial and human resources. Figure 7 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 rather 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. This adds a new dimension to the digital divide: a divide among the haves and have-nots in terms of digital service capabilities, which are crucial for Big Data capacities. It makes a critical difference if 1 in 50 or 1 in 500 of the national workforce is specialized in software and computer services (e.g. see Finland vs. Mexico in Figure 7).

Figure 7: 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.

Data as a commodity: in-house vs. outsourcing

There are two basic options on how to obtain such Big Data services: in-house or through outsourcing. On the firm-level, Brynjolfsson, Hitt, and Kim (2011) find that data driven decision making is slightly stronger correlated with the presence of an in-house team and employees than with general ICT budgets, which would enable to obtain outsourced services. This suggests that in-house capability is the stronger driver of organizational change toward Big Data adoption. The pioneering examples of large in-house Big Data solutions include loyalty programs of retailers (e.g. Tesco), tailored marketing (e.g. Amazon), or vendor-managed inventories (e.g. Wal-Mart). However, those in-house solutions are also notoriously costly.

Outsourcing solutions benefit from the particular cost structure of digital data, which have extremely high fix-costs and minimal variable costs (Shapiro and Varian, 1998): it might cost millions of dollars to create a database, but running different kinds of analysis is comparatively cheap, resulting in large economies of scale for each additional analysis. This economic incentive leads to an increasing agglomeration of digital data capacities in the hands of specialized data service provider which provide analytic services to ad hoc users. For example, specialized Big Data provider companies provide news reporters with the chance to consult the historic voting behavior of senators, restaurants with the intelligence to evaluate customer comments on the social ratings site Yelp, and expanding franchise chains with information on the vicinity of gas stations, traffic points or potential competition in order to optimize the placement of an additional franchise location (Hardy, 2012a). Others specialize on on-demand global trade and logistics data, which include on the contracting, packing and scanning of freight, documentation and customs, and global supply chain finance (Hardy, 2012b) and again others offer insights from Twitter and other social networking sites. Being aware of the competitive advantage of having in-house knowledge of Big Data Analysis, but also about the sporadic need to obtain data that is much more cost-effectively harnessed by some third party provider, many organizations opt for a hybrid solution and use on-demand cloud resources to supplement in-house deployments (Dumbill, 2012).

In this sense, data itself becomes a commodity and therefore subject to existing economic divides. With an overall revenue of an estimated US$ 5 billion in 2012 and US$ 10 billion in 2013 globally (Feinleib, 2012), the Big Data market is quickly getting bigger than the size of half of the world's national economies. Creating an in-house capacity or buying the privilege of access for a fee "produces considerable unevenness in the system: those with money – or those inside the company – can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access" (boyd and Crawford, 2012; p. 673-674). The existing unevenness in terms of economic resources leads to an uneven playing field in this new analytic divide.

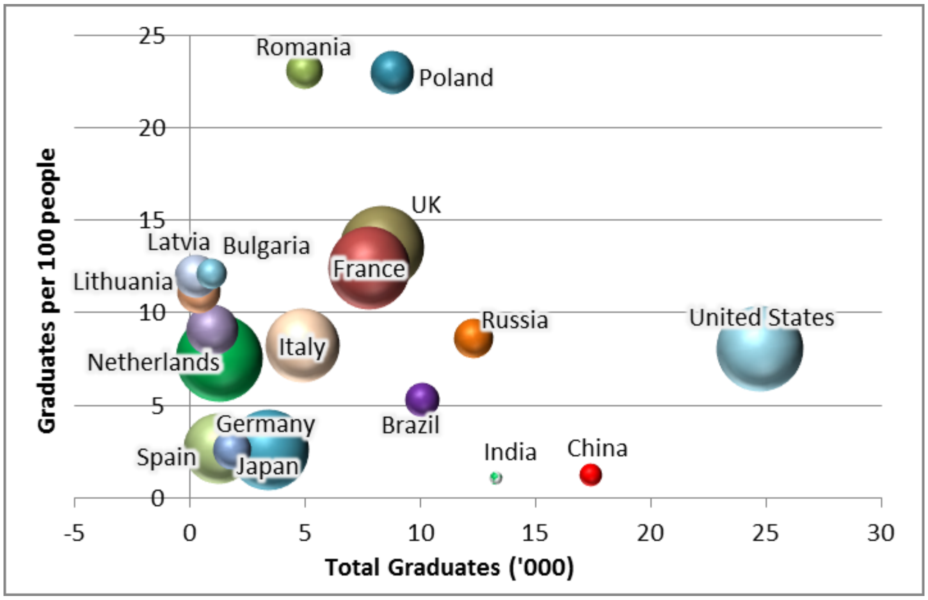
Capacities & Skills

Additional to supporting hardware and service capabilities, the exploitation of Big Data also requires data-savvy managers and analysts and deep analytical talent (Letouzé, 2011; p. 26 ff), as well as capabilities in machine learning and computer science. Hal Varian, chief economist at Google and Professor emeritus at the University of California at Berkeley, notoriously predicts that "the sexy job in the next 10 years will be statisticians… And I'm not kidding" (Lohr, 2009;

p.1). Statisticians and awareness about the importance of statistical capabilities are rare in developing countries. In a characteristics example, Ghana's statistical authorities took 17 years to adopt the UN system of national accounts from 1993. After up-dating their method in 2010 the surprised statisticians found that Ghana's GDP was 62 % higher than previously thought (Devarajan, 2011). 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). First 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).

Figure 8 shows that the perspectives in this regard are actually mixed for different parts of the developing world. Some developing countries with relatively low income levels achieve extremely high graduation rates for professionals with deep analytical skills (see e.g. Romania and Poland, which are high up on the vertical y-axis in Figure 8). In general, countries from the former Soviet bloc (also Latvia, Lithuania, Bulgaria, and Russia) produce a high level of analysts. Other countries, such as China, India, Brazil and Russia produce a large number of analysts (far to the right on the x-axis in Figure 8, which mainly relates to their population size in absolute terms). In 2008, these so-called BRIC countries (Brazil, Russia, India and China) produced almost 40 % of the global professionals with deep analytical skills, twice and many as the United States. Traditional power-houses of the global economy, such as Germany and Japan, are comparatively ill-prepared for the human skills required in a Big Data age. This leads to the long-standing and persistent discussion about brain drain, and of the eventual possibility that professionals from developing countries will build bridges of capabilities between developed and developing countries (Saxenian, 2007).

***Figure 8: 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).

One way of dealing with the shortage and fostering the creation of skilled professionals are collective data analysis schemes, either through collaboration or competition. This does not only apply to developing countries. A survey of leading scientists from the Journal *Science* 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). Wikis to collectively decode genes or analyze molecular structures have sprung up over recent years (Waldrop, 2008), and specialized platforms of distributed human computing 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 a dozen of data analysis competitions with cash prizes between US$ 150 and US$ 3,000,000 (Carpenter, 2011). In one competition, a Ph.D. student in glacier mapping outperformed NASA's longstanding algorithms to measure the shape of galaxies (Hardy, 2012b). In another example, 57 teams (including from Chile, Antigua and Barbuda, and Serbia) helped an Australian statistician to predict the amount of money spend by tourists (a value insight for a mere US$ 500 cash price) (Hyndman, 2010).

Incentives: positive feedback

The third side of the conceptual framework from Figure 1 represents the social construction of technological change through policy strategies that aim at chosen normative aspects of development. One way of doing this is to positively encourage and foster desired outcomes.

Financial incentives and subsidies

As so often, money is not the sole solution, but it makes things easier. One concrete example of government subsidies is the Office of Cyberinfrastructure (OCI) of the U.S. National Science Foundation (NSF), which counts with a budget US$ 700 and US$ 800 million to invest, among other objectives, 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 includes fostering data visualization (Frankel and Reid, 2008) (see Figure 9 for a simple example). NSF and the academic Journal *Science* have hosted a data visualization competition for nine consecutive years (Norman, 2012).

Figure 9: 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/6170795/BigDataArticle

A much more resource intensive effort, also from the U.S., refers to the approximately US$ 19 billion of the American Recovery and Reinvestment Act that is earmarked to encourage physicians to adopt electronic medical recordkeeping systems (Bollier, 2010). Digitizing recordkeeping makes health care more versatile and contributes to important savings. It is estimated that a correction of the abnormalities in Figure 3b (deviate patterns of Medicare spending) would save up to 33 % of total health care spending in the U.S. (Darthmouth, 2012), which represents 16 % of U.S.'s GDP. In developing countries, health care expenditure ranges between 4 – 8 % of GDP.

Exploiting public data

Another incentive for Big Data consists in the exploitation of the natural quasi-monopoly held by the public sector for many areas of social data (Kum, Ahalt and Carsey, 2011; WEF and Vital Wave, 2012). 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 "public", around one quarter of the available data resources could be liberated for Big Data Analysis.

The ongoing discussion about the openness of digital government data moves along two main axes (Figure 10a). One is rather technical and refers to the accessibility of the data format. Information listed in PDF files, for example, are less accessible and actionable than data published in structured Excel spreadsheets, while those proprietary spreadsheets are again less accessible than open source databases like CSV. The emerging gold standard are so-called "linked data" (Berners-Lee, 2006), which refers to datasets that are described by an open standard metadata layers (such as uniform resource identifier (URI) and resource description framework (RDF)) that makes the data readily readable and sortable for humans and machines. The other axis refers to the kind of data source. The most straightforward kind of source is traditional public statistics, such as produced by household surveys and censuses. Geospatial data[6] is also among the most widely published public data and is useful for a large amount of applications, as previously discussed. Public procurement and expenditure data is often less transparent, even so some developing and developed governments have made important advances (e.g. Suárez and Laguado, 2007; or usaspending.gov). The borderline between usefulness and legality is often in question in the case of publishing vast amounts of

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

(sometimes classified) documents through portals like Wikileaks (e.g. Sifry, 2011). Here the topic of Big Data for intelligent decision-making intersects with the more contentious topic of public transparency and State secrecy, which often runs under the heading of "open governments" (Lathrop and Ruma, 2010; Concha and Naser, 2012). The numbers 1-4 in Figure 10a loosely classify the perceived level of challenges encountered when publishing open government data.[7]

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, 2011)[8]. Several dozen developing countries have set up 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 local housing, the condition of highways, and the location of public bicycle stands. Data.gc.ca from Canada and Data.gov from the U.S. stand out with over 260,000 and 370,000 raw and geospatial datasets from a couple of hundred agencies respectively.[6] On the one hand, the open access model allows everybody to access this wealth of data collected and published by the most advanced countries. This provides important opportunities for developing countries, such as shown by the case of the usefulness of weather and climate data (GFDRR, 2012). On the other hand, data about housing, geography, traffic, and health is certainly most useful to the host country. In the case of France, 76 % of the data is national, 12 % regional, 10 % local and departmental, and only 2 % international (Vincey, 2012). Therefore, local data production capacity still provides an international development advantage.

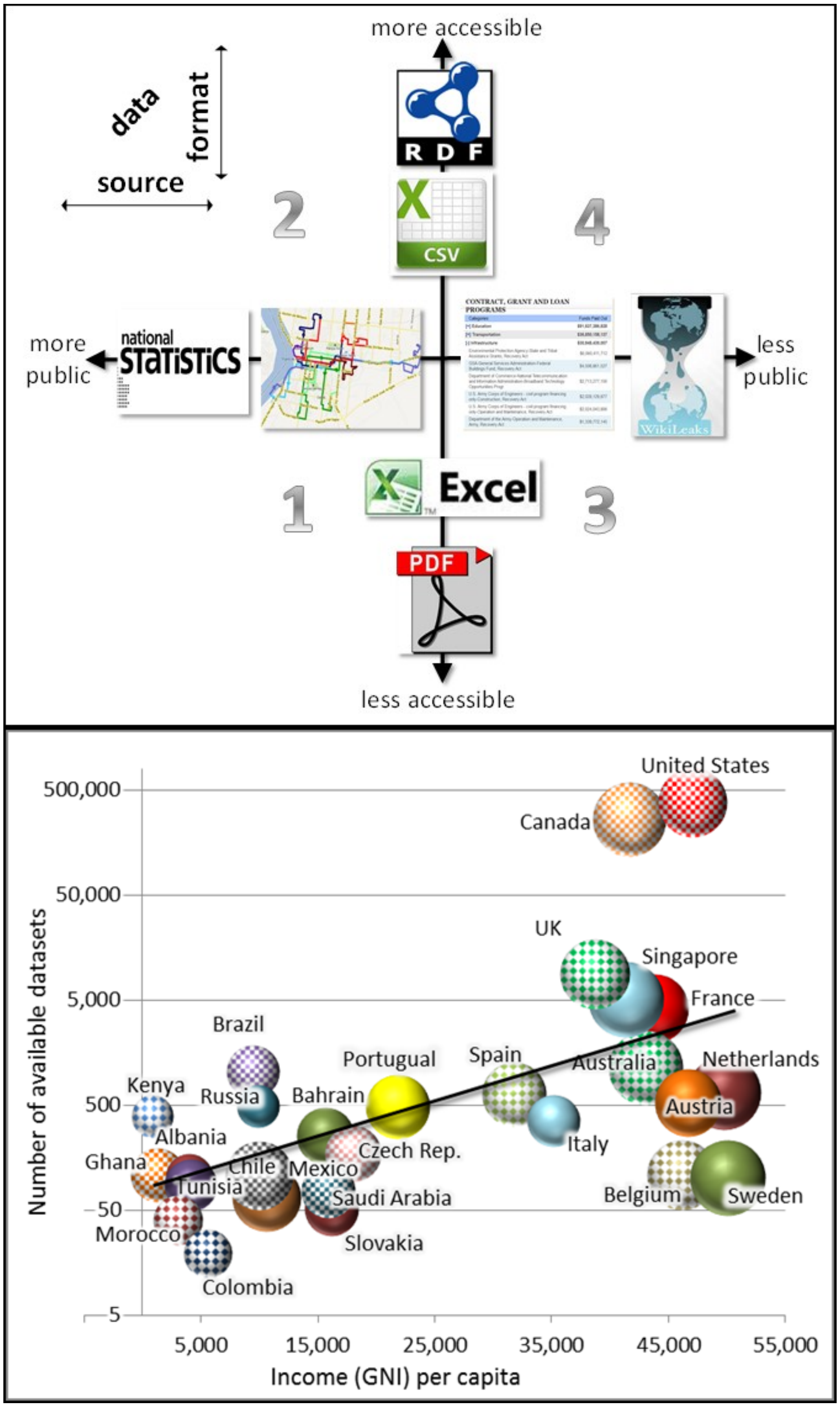
The good news is that an open data policy does not seem to be strongly correlated with the level of development of the country. Figure 10b shows that the number of databases provided on these central government portals correlate only weakly with the economic wellbeing of the country (horizontal x-axis) and the perceived path trajectory of transparency in the national public sector (the size of the bubbles presents the most widely used index of perceived

[7] Based on the collective sentiment of the experts that participated in the workshop: "Open Data 4 Development (OD4D): Datos abiertos para una economía del conocimiento más inclusiva" (Jan. 07, 2013; United Nations ECLAC, Santiago, Chile; http://www.od4d.org).

[8] At the end of 2012, 55 government around the world have signed the Open Government Declaration from which this quote is taken.

transparency and corruption worldwide; Transparency International, 2011). On average, those governments of our sample with more than 500 publicly available databases on their open data online portals have 2.5 times the per capita income, and 1.5 times more perceived transparency than their counterparts with less than 500 public databases. Notwithstanding, Figure 10 also shows that several governments from developing countries are more active than their developed counterparts in making databases publicly available (see e.g. Kenya, Russia and Brazil).

***Figure 10: Open Government datas:** (a) schematic conceptual framework; (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 bubbles, more transparent) (year=2011; n=27).*



Source: own elaboration, (b) based on the 27 official open data portals; World Bank, 2010; and Revenue Watch Institute and Transparency International, 2011. Note: First launched in 1995, 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 and often hidden activity, subjective perceptions turn out to be the most reliable method: www.transparency.org/research/cpi)

Regulation: negative feedback

The other kind of tools to guide the Big Data paradigm into the desired development direction consists in the creation of regulations and legislative frameworks. This touches on many of the longstanding issues that have been discussed for years in the ICT-community (e.g. Lessig, 2000). It involves security (e.g. how frequent is data theft and espionage?), intellectual property (e.g. who owns which data, who owns which data analysis results, and is a detected data pattern patentable?), liability (who is responsible for inaccurate data that leads to negative consequences?), and interoperability (who defines the standards to enable data exchange, and are they open or proprietary?).

Control and privacy

Concerns about privacy and State and corporate control 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). As early as 1948, some 25-30 years before scholars like Bell (1973) and Masuda (1980) started to talk about the "Information Age", George Orwell described a rather terrifying vision of the Information Society: "By comparison with that existing today, all the tyrannies of the past were half-hearted and inefficient" (Orwell, 1948; 2, 9). Fact of the matter is that "any data on human subjects inevitably raise privacy issues" (Nature Editorial, 2007; p. 637). Digital information always leaves a potential trace that can be tracked and analyzed (Andrews, 2012). One distinction that is often made in the Big Data discussion is whether or not the tracked data is generated actively or passively, and voluntarily or involuntarily (King, 2011). For example, the collection of Big Data on social activity often blurs the difference between being in public (i.e. sitting in a park) and being public (i.e. actively courting attention) (boyd & Marwick 2011). Traditional research surveys are an example of active and voluntary data provision. In the United States, the Food and Drug Administration (FDA) and Department of Health and Human Services have passed regulations that have empowered so-called Institutional Review Boards (IRBs) to approve, require modifications in planned research prior to approval, or disapprove research involving humans. Such IRBs approval processes have become a standard part of a graduate education at American research universities and "scientists must meet strict rules on any research on human subjects. In contrast, private firms are largely free from such constraints, and already have wide latitude to snoop on, and data mine, their employees' work habits". (Buttler, 2007; p. 645). These issues are much less regulated in developing countries, be it in the private sector or academia. A less regulated example of active and voluntary data provision refers to online user ratings of products or services, such as customer reviews or scaled ratings. These sources are frequently used for large-scale data analysis. An example of voluntary passive data provision is when users knowingly allow online retailers and search engines to personalize shopping recommendations and search results based on passed interactions with the system. An even more contentious example of involuntary passive data provision is the tracking of Twitter comments or mobile phone locations (Andrews, 2012).

While the fine-tuning of intelligent search mechanisms and the personalization of shopping experiences are seen desirable by many users, the issue of privacy becomes especially delicate when personalized data is used for control. Orwell (1948; 1,3) warned especially about the manipulation of democratic processes through personalized control and brainwashing. In present times, the analysis of various kinds of Big Data (including credit card repositories) have resulted in well-known concepts as "Soccer Moms", "America's Home-Schooled" or "Late-Breaking Gays" (Penn and Zalesne, 2007), which have become decisive swing groups in American party pooling for votes in democratic elections. In the best case scenario, the identification of these groups enables a political candidate for democratic office to spin a message to please an identified group of interest. The result is populism and not the democratic representation of the people through a free mandate, such as foreseen in most democratic constitutions (Hilbert, 2007; Ch. 2.3). In the worst case, the political candidate uses this information to spin a message to manipulate the identified group. The pinpointed manipulation of citizens evidently already moves into the direction of Orwellian brainwashing.

The democratic flipside to the transparent citizen is the transparent State, which returns to the discussion of "open government data", this time not from the perspective of voluntary projects, but from the perspective of mandatory regulation. 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).

Interoperability of isolated data silos

One of the main challenges of harnessing Big Data consists in bringing data from different sources together. Large parts of valuable data lurk in "data silos" of different departments, regional offices, and specialized agencies. Fragmentation impedes the massive and timely exploitation of data. Manyika, et al. (2011) show that the data landscape in sectors like education and health tends to be more fragmented than the rather concerted data landscape of banking or insurance services, whose databases speak the same informatics language. Data interoperability standards are becoming a pressing issue for the Big Data paradigm in both developed (NSF, 2007), as well as in developing countries: several years ago, Latin American governments have started to work on a White Book on e-Government interoperability in Latin America (UN-ECLAC, 2007), but over recent years, the topic as a whole has lost momentum in the region (de la Fuente, 2012).

Critical reflection: all power to the algorithms?

We end this article with a critical reflection on the broader implications of the Big Data paradigm for development. Placing computer-mediated analytic treatment of data at the forefront of decision-making also implies the encouragement of machinated decision-making over human evaluation. In the past, the vast majority of information processing was executed by managers, analysts, and human data crunchers (Nelson, 2008)[9]. Human evaluators have been overtaken by machines in many fields. By now, Big Data based artificial intelligence diagnosis tools that detect aneurysms have a success rate of 95% versus 70 % for human radiologists (Raihan, 2010). 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. These processes might be directly dictated by a human being or by another algorithm (such as an evolutionary algorithm), which again was dictated by a software specialist. Unfortunately, the programmer rarely is able to consider all the intricate complexities of a constantly evolving environment, which consists of a large number of interdependent parts, which pursue different goals. While some of the results are rather amusing (such as a book on flies that was offered for US$ 23 million on Amazon by competing algorithms that calculated supply and demand patterns, Slavin, 2011), while others can have disastrous consequences that affect the stability of entire economies, such as shown by the example of "black-box" trading (or algorithmic trading). From a starting point near zero in the

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

mid-1990s, algorithmic trading is 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).

The common reason for the imperfect nature of algorithms is that fact that most current algorithms are mainly informed by the world as it was or, at best, as it is. Fed by a large number of past experiences, common algorithms can predict future development if the future is similar to the past. In order to do so, it is not even necessary to be able to explain the ongoing dynamics of the past. For example, a social networking site like Facebook or Twitter might not be able to answer the more fundamental questions like "why are people saying what they are saying?" and "why are people behaving like they are behaving?" but they can tell us that they presently do, and, if nothing changes, that they will continue to do in the future. "Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day" (Anderson, 2008, p.1). In other words, Google can predict without explaining, nor understanding, but simply by looking for patterns from the past.[10] Since explaining and predicting are notoriously different (Simon, 2002; Shmueli, 2010), blind prediction algorithms can disgracefully fail if the environment evolves, since the insights are based on the past, not on a general understanding of the overall dynamics. Considering the exponential complexity arising from mutual endogenous and exogenous influences among stock market traders, trading algorithms, and the general economic environment, it is not surprising that specialized trading algorithms are not able to handle all cases of a quickly changing trading landscape.

Another consequence is that algorithms based on data from the past will naturally reinforce past behavior. Garland (2012) reports that many corporate and government leaders got used to hearing reports that confirm data patterns that they are used to seeing, and react with "confusion, anger, and psychological transference" when confronted with future scenarios that are discontinuous of existing patterns. He concludes that data pattern based decision-making makes it actually "harder for us all to adapt to a changing world" (p. 1). The repeated confrontation with personal past behavior not only leads to cognitive dissonance, but potentially also to social conflict. For example, personalized online search machines use algorithms to selectively guess what information a user would like to see based on past

[10] The father of the Minimum Description Length, Jorma Rissanen (2010) defends the Platonic and Kantian point that we will never be able to perceive reality as it is and that therefore "one can never find the 'true' data-generating distribution" (p. 2). Therefore, "instead of trying … to get a model which is close to the 'true', and in fact nonexistent, target distribution, the objective is to extract all the useful and learnable information from the data that can be extracted with a model class suggested" (p. 6), which then can give accurate predictions.

information about the user (such as location, past clicking behavior and search history), which is useful if the user would like to fine-tune research results or shopping suggestions. However, as a result, the algorithms tend to show only information which agrees with the user's past viewpoint. For example, Pariser (2011, p. 9) reports two different people performing an online search for "BP". While one got investment news about British Petroleum, another got information about the Deepwater Horizon oil spill. The constant reconfirmation of personal viewpoints can easily lead to polarization and extremism. Polarization is one of the innate enemies of the democratic process of creating one common will of the people through critical reflection of alternative viewpoints (Habermas, 2000; Hilbert, 2007, Ch. 2.1).

It is important to underline that algorithms do not necessarily have to be based purely on Big Data sets that explain past behavior. Agent-based model, for example, are increasingly getting better in predicting the outcome of social complexities of even unknown future scenarios through computer simulations that are based on a collection of mutually interdepend algorithms. Some hope that the combination of data from the past and computational modeling of future scenarios will help us to get a better understanding of ongoing social complexities (e.g. Farmer and Foley, 2009).

Conclusion

Recently, much has been written and discussed about the Big Data paradigm. A systematic review of over 100 pieces of mainly recent literature and several pieces of hard fact empirical evidence show that the Big Data paradigm holds both promises and perils for development dynamics. 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 during the diffusion process through social networks (Rogers, 2003). As with all previous examples of technology-based innovation for development, also the Big Data paradigm runs through a slow and unequal diffusion process that is compromised by 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. This divide does not only refer to the availability of information, but to intelligent decision-making and therefore to 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. However, in a world where we desperately need further insights into development dynamics, Big Data Analysis can be an important tool to contribute to our understanding of and improve our contributions to manifold development challenges.

References

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

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, (Science: Discoveries). Retrieved from 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.

Bateson, G. (2000). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University of Chicago Press.

Bell, D. (1973). *The Coming of Post-Industrial Society: A Venture in Social Forecasting*. New York, NY: Basic Books.

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 Trans parency and Accountability Initiative. Retrieved from http://www.revenuewatch.org/sites/default/files/Russia_TAI_eng.pdf

Beniger, J. (1986). *The Control Revolution: Technological and Economic Origins of the Information Society*. Harvard University Press.

Berners-Lee, T. (2006, July 27). Linked Data. *W3 Design Issues*. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

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

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 and 2011 IEEE Third International Confernece on Social Computing (SocialCom)* (pp. 157–164). Presented at the Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Confernece on Social Computing (SocialCom), IEEE. doi:10.1109/PASSAT/SocialCom.2011.122

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). Retrieved from https://www.mckinseyquarterly.com/Are_you_ready_for_the_era_of_big_data_2864

Brynjolfsson, E., & Hitt, L. M. (1995). Beyond Computation: Information Technology, Organizational Transformation and Business Performance. *Journal of Economic Perspectives*, *14*, 23–48.

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.

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

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

Devarajan, S. (2011). Africa's statistical tragedy. *Africa Can...end Poverty*. Retrieved from 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). Retrieved from http://scholarworks.umass.edu/cpo/vol1/iss1/4

Dumbill, E. (2012, January 11). What is big data? An introduction to the big data landscape. *O'Reilly Radar*. Retrieved from 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 Trans parency and Accountability Initiative. Retrieved from http://www.revenuewatch.org/sites/default/files/India_TAI_eng.pdf

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. Retrieved from http://www.slideshare.net/bigdatalandscape/big-data-trends

FOI (Freedom of Information). (2012). Freedom of information legislation - Wikipedia, the free encyclopedia. *Wikipedia*. 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

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

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

Garland, E. (2012, April 5). Peak Intel: How So-Called Strategic Intelligence Actually Makes Us Dumber. *The Atlantic*. Retrieved from http://www.theatlantic.com/international/archive/2012/04/peak-intel-how-so-called-strategic-intelligence-actually-makes-us-dumber/255413/#.T4BS2naAlR8.mailto

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). *The GFDRR Labs*. Retrieved from https://www.gfdrr.org/opendri

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

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

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 Trans parency and Accountability Initiative. Retrieved from http://www.revenuewatch.org/sites/default/files/Philippines_TAI.pdf

Guerin, P. J., Bates, S. J., & Sibley, C. H. (2009). Global resistance surveillance: ensuring antimalarial efficacy in the future. *Current Opinion in Infectious Diseases*, *22*(6), 593–600. doi:10.1097/QCO.0b013e328332c4a7

Habermas, J. (2000). *The Inclusion of the Other: Studies in Political Theory* (reprint.). The MIT Press.

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

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

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. (2007). *Digital Processes and Democratic Theory: Dynamics, Risks and Opportunities that Arise when Democratic Institutions Meet Digital Information and Communication Technologies*. peer-reviewed online publication; Google Books. Retrieved from http://www.martinhilbert.net/democracy.html

Hilbert, M. (2011). 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. (2012). Towards a Conceptual Framework for ICT for Development: Lessons Learned from the Latin American "Cube Framework". *Information Technologies & International Development*, *8*(Winter; Special issue: ICT4D in Latin America), forthcoming.

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 Trans parency and Accountability Initiative. Retrieved from http://www.revenuewatch.org/sites/default/files/SouthAfrica_TAI.pdf

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

IBM. (2011). *Vestas: Turning climate into capital with big data* (Case study). Retrieved from 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

ITU (International Telecommunication Union). (2011). *Measuring the Information Society 2011*. Geneva: International Telecommunication Union, ITU-D. Retrieved from http://www.itu.int/publ/D-IND-ICTOI-2011/en

Jorgenson, D. W. (2002). *Econometrics, Vol. 3: Economic Growth in the Information Age* (1st ed.). The MIT Press.

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

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

Lohr, S. (2009, August 6). For Today’s Graduate, Just One Word: Statistics. *The New York Times*. Retrieved from 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). Retrieved from http://www.martinhilbert.net/WorldInfoCapacity.html

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. Retrieved from 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. Retrieved from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

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

Melville, N., Kraemer, K., & Gurbaxani, V. (2004). Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value. *MIS Quarterly*, *28*(2), 283–322.

Moreno, R. (2012). *Mexico: A Moment of Opportunity. Increasing Transparency and Accountability in the Extractive Industries* (Working Paper Series 2012). Revenue Watch Institute and Trans parency and Accountability Initiative. Retrieved from http://www.revenuewatch.org/sites/default/files/Mexico_TAI_eng.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. 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). Community-based Data Interoperability Networks (INTEROP). *Office of Cyberinfrastructure (OCI)*. Retrieved March 28, 2012, from http://www.nsf.gov/od/oci/about.jsp

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

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

Open Government Partnership. (2011, September). Open Government Declaration. Retrieved from http://www.opengovpartnership.org/sites/www.opengovpartnership.org/files/page_files/OGP_Declaration.pdf

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

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin.

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. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2880/3264

Penn, M., & Zalesne, E. K. (2007). *Microtrends: The Small Forces Behind Tomorrow's Big Changes*. Twelve.

Peres, W., & Hilbert, M. (2010). *Information Societies in Latin America and the Caribbean Development of Technologies and Technologies for Development*. Santiago: United Nations ECLAC. Retrieved from 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. Retrieved from http://www.carlotaperez.org/papers/basic-technologicalrevolutionsparadigm.htm

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

Raihan, I. (2010). Managing Big data. Retrieved from 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. Retrieved from 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.

Rogers, E. M. (2003). *Diffusion of Innovations, 5th Edition*. New York: Free Press.

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. Retrieved from 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

Sen, A. (2000). *Development as Freedom* (Reprint.). New York: Anchor.

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

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

Sifry, M. L. (2011). *WikiLeaks and the Age of Transparency*. OR Books.

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. Retrieved from http://dx.doi.org/10.1017/CBO9780511493164

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

Suárez Beltrán, G., & Laguado Giraldo, R. (2007). *Manual de contratación pública electrónica para América Latina. Bases conceptuales, modelo legal, indicadores, parámetros de interoperabilidad* (No. 20). Santiago Chile: United Nations ECLAC.

Sunstein, C. (2003). The Law of Group Polarization. In J. Fishkin & P. Laslett (Eds.), *Debating Deliberative Democracy* (1st ed., pp. 80–102). Wiley-Blackwell.

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

Transparency International. (2011). *Corruption Percpetion Index 2011*. Retrieved from 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

UNU (United Nations University). (2012). *Rainfall Variability, Food Security and Human Mobility: an approach for generating empirical evidence* (No. No. 10). Bonn: United Nations University Institute for Environment and Human Security (UNU-EHS). Retrieved from http://www.ehs.unu.edu/file/get/9921.pdf

Valente, T. W., & Saba, W. P. (1998). Mass media and interpersonal influence in a reproductive health communication campaign in Bolivia. *Communication Research*, *25*(1), 96.

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

WEF (World Economic Forum), & Vital Wave Consulting. (2012). Big Data, Big Impact: New Possibilities for International Development. *World Economic Forum*. 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, John 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

Wolfe, N., Gunasekara, L., & Bogue, Z. (2011, February 2). Crunching digital data can help the world. *CNN*. Retrieved from http://articles.cnn.com/2011-02-02/opinion/wolfe.gunasekara.bogue.data_1_personal-data-big-data-bushmeat?_s=PM:OPINION

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. Retrieved from 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