Title
e-Science for Digital Development: ICT4ICT4D

Author
Hilbert, Martin

Publication Date
2015

Peer reviewed

Development Informatics

Working Paper Series

The Development Informatics working paper series discusses the broad issues surrounding information, knowledge, information systems, and information and communication technologies in the process of socio-economic development

Paper No. 60

e-Science for Digital Development: *"ICT4ICT4D"*

MARTIN HILBERT

2015

ISBN: 978-1-905469-54-3

Published by: ***Centre for Development Informatics***
Institute for Development Policy and Management, SEED
University of Manchester, Arthur Lewis Building, Manchester, M13 9PL, UK
Email: cdi@manchester.ac.uk Web: http://www.cdi.manchester.ac.uk

View/Download from:
http://www.seed.manchester.ac.uk/subjects/idpm/research/publications/wp/di/

Educators' Guide from:
http://www.seed.manchester.ac.uk/subjects/idpm/research/publications/wp/di/educdi/

Table of Contents

e-Science for Digital Development: *"ICT4ICT4D"*

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

2015

Abstract

While the ICT for development (ICT4D) community is well aware about the far-reaching changes introduced by the digital age, it is remarkably slowly getting used to the idea that digital tools also revolutionize its very own core business: research. Information and communication technology (ICT) is currently transforming the way knowledge is created and insights are obtained. This applies to inductive empirical inquiry (especially 'big data'), as well as deductive theoretical scholarship (especially 'computer simulation modeling').

This paper explores best practices of the application of such e-science in the field of development and contributes an informed perspective to intensify an outstanding discussion within the ICT4D community. ICT should not only be used for development (ICT4D), but also for the constant updating of our understanding of digital development, in order to fine-tune policies and project designs: 'ICT4ICT4D'.

On the one hand, the paper shows that the consideration of this double role of ICT has the potential to significantly increase the impact of ICT4D. The digital 'big data' footprint provides unprecedented insights into dynamics in the development context that have traditionally been lacking empirical evidence; and the modular nature of computer simulations allows us to study scenarios for specific contexts, which lessens the dependence on one-size-fits-all models. This increases potential impact for ICT4D because both of them provide complementary tools to assure that the application of ICT for development purposes is rooted in their context-dependent local reality.

On the other hand, developing countries and the ICT4D community face important challenges when applying these tools, which—as all technological innovations—should never be adopted uncritically.

A. Two Digital Research Legs: Empirical Induction and Deductive Models

Digital information and communication technologies (ICTs) are currently revolutionizing the way research is carried out. This affects both main components of any scientific project: empirical work with data (the main driver of induction), and theoretical model building (the main driver of deduction).

Given that ever more of human conduct is taking place in digital networks, and given that digital conduct inevitably leaves a digital footprint, the social sciences currently have access to an unprecedented amount of data on the most diverse aspects of the social fabric and its development dynamics (Lazer et al., 2009). The catch-phrase here became 'big data' (Mayer-Schönberger and Cukier, 2013; Manyika et al., 2011; Kolb and Kolb, 2013; Letouzé, 2012; WEF, 2012; Hilbert, 2015). Its impact on the social sciences has been compared with the impact of the invention of the telescope for astronomy and the invention of the microscope for biology (providing an unprecedented level of detail about the system of interest). Confronted with such increase in the level of perceivable granularity in social dynamics, social scientists have an inevitable obligation to make use of it to inform analysis, policy and project design. Since ICT for development (ICT4D) unavoidably involves ICT, most ICT4D applications automatically produce such a digital footprint.

For practical purposes, it is important to recognize that the crux of the big data paradigm is twofold. For one, it refers to new sources of data. The digital footprint created with each digital communication and transaction can replace traditional data sources (like surveys) with proxy indicators that correlate with the variable of interest. The benefit is the low cost and real-time availability of the digital proxy indicator. The epitome is Google's illustrious use of 50 million most-common search terms as a proxy for the spread of the seasonal flu (Ginsberg, et al., 2009). Secondly the notion of big data goes beyond data itself, and focuses on methods of data analytics to inform intelligent decisions. Independent from the specific giga-, tera-, peta-, or exabyte scale, the big data paradigm argues to systematically place analytic treatment of data 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" to "Knowledge Societies" (Hilbert, 2013). Building on the digital infrastructure that led to vast increases in information during recent decades, the big data paradigm focuses on converting this digital information into knowledge that informs intelligent decisions. Continuing with the previous example, Google processed an impressive 450 million different mathematical models in order to test for correlations between online search terms and flu outbreaks reported by official data. Eventually, 45 search terms were identified that outperformed traditional models of flu outbreak with real-time predictions (Ginsberg, et al., 2009).

While the opportunities of big data are enormous, especially for developing countries in which traditional statistics are scarce (Letouzé, 2012; WEF, 2012; Hilbert, 2015), the very same Google flu trend case also exemplifies the ultimate limitation of (big) data analytics: the fact that all data is from the past. The application of the same Google flu trend algorithm became increasingly out of sync with actual flu epidemics over time (Lazer et al., 2014). The reason is straightforward: reality (search behaviour) is changing over time, which provides different input and changes the results if the same algorithm is used. The best big data analytics can do is to update estimations in 'real-time': so called 'nowcasting' (Giannone, Reichlin and Small, 2008; Carrière-Swallow and Labbé, 2013) (both the terms are actually misnomers, since the very act of recording converts it into 'past-time' data). Since data is always from the past, it can only detect patterns that have occurred in the past. When the past and the future follow the same stationary logic, data analytics is extremely useful to predict future patterns. However, if significant changes occur in the system's dynamic, empirical statistics are at best limited, if not deceiving.

This argument is known as the 'Lucas critique' in economics (Lucas, 1976), as 'Goodhart's law' in finance (Goodhart, 1976) and as 'Campbell's law' in education (Campbell, 1976), all dating back to 1976. Nobel Prize-winning economist Robert Lucas criticized colleagues who used sophisticated statistics to make economic predictions ('econometrics') in order to inform policy making. He argued that no useful information can emerge from such analysis because "any change in policy will systematically alter the structure of econometric models" (Lucas, 1976). This is crucial for ICT4D. Development work has the explicit goal to create a future that is significantly different from the past. It explicitly aims at changing aspects of the *modus operandi* of the system. Considering the diversity of development settings, the outcome is very context dependent and almost always unique. A developing Africa is not simply an extrapolated version of Europe's past development trajectory. This limits the usefulness of data from a specific case of the past.

In order to predict a future that has never been, theory-driven models are necessary. These allow variables to be adjusted with values that have never existed in statistically observable reality. ICT also acts as a game changer in this challenge. Computational simulations allow to set up theory-driven models that greatly expand the scope and level of sophistication of traditional 'paper-and-pen' models. While traditional models are only able to handle a very limited number of variables (at most a dozen or so), today's computational power allows creating mathematically formalized models with thousands and even millions of dynamic variables. Such computer simulations of artificial societies have no conceptual limitations on the achievable level of detail and accuracy. Most recent simulations are based on individual agents ('agent-based models'), resulting in an emergent interplay between bottom-up and top-down dynamics (Schelling, 2006; Epstein and Axtell, 1996; Miller and Page, 2007; Wilenksy and Rand, 2015).

The combination of both empirical (big) data analysis of the digital footprint, and computer simulation models of possible scenarios are understood as e-science in this paper (compare with related uses of the term in Hummon and Fararo, 1995; Lazer et al., 2009). Sections 2 and 3 review the characteristics of these two aspects of e-science for development, while section 4 turns to the ensuing opportunities and challenges. The review draws from over 100 referenced articles. Those are taken from a large variety of sources and domains, but all inform the argument of how ICT can be used in ICT4D. As such, the contribution of the paper consists in the collection and analysis of previously diffuse evidence about the potential of two concrete components of modern research, big data and computer simulations, for ICT4D.

B. Big Data for Development

The value unleashed by big data to inform decision has been referred to as "the new oil" (Kolb and Kolb, 2013) and recent literature has started to point to the important opportunities that big data opens up for development (Letouzé, 2012; WEF, 2012; Hilbert, 2015). The OECD is convinced that "big data now represents a core economic asset that can create significant competitive advantage" (OECD, 2013); the United Nations argues that "new technologies are…creating unprecedented possibilities for informing and transforming society and protecting the environment… faster and more detailed than ever before" (IEAG, 2014); and the UN Economic and Social Council has already reported to the UN Secretary General that "big data have the potential to produce more relevant and more timely statistics than traditional sources of official statistics, such as survey and administrative data sources" (UN Statistical Commission, 2014).

B1. Characteristics of Big Data

The big data paradigm can be characterized by some five general features (Hilbert, 2015). Big data:

- replaces random sampling with the ambition to *capture all* there is (sampling n = universe N).
- is often accessible in *real-time*.
- is produced anyways as a low-cost and almost inevitable *digital footprint*.
- is messy and incomplete, which can be compensated by data redundancy from different sources, often called *data fusion*.
- uses exploratory data mining and *machine-learning tools*, which replace the need for theory with plain pattern detection, and therefore leads to a theory-free and automated way to inform the decision-making process.

In related literature, the first three are often referred to as Volume, Velocity and Variety (Manyika, et al., 2011; WEF, 2012).

B2. Big Data and Development

We will review concepts and applied examples of each of these five characteristics from the field of development.

B2.1. Universal sampling. One of the biggest potentials for big data in developing countries consists in mobile phone data. With a global penetration of over 95% (ITU, 2014) and around 75% access among those making US$1 per day or less (Naef et al., 2014), mobile phones became an important source of social data in developing countries. Their pervasiveness allows to access important features of society as a whole, without the need to set up sophisticated sampling mechanisms. The key enabler is the detection of the correlations between mobile phone usage and some other 'real-world' conduct (Raento, Oulasvirta and Eagle, 2009). On the one hand, understanding mobile phone usage provides insights into the nature of technology diffusion and the digital divide (Blumenstock, Eagle, and Fafchamps, 2012), as well as commercial telecom operators with "critical information for the personalization and adaptation of mobile-based services to the behavioral segments identified" (Frias-Martinez and Virseda, 2013; p. 36). In other words, patterns in human behaviour are used to influence mobile phone development. This might often even be the incentive for such studies. On the other hand, mobile phone records can also be used to infer demographic, socio-economic and other behavioral traits based on the continuously registered digital conduct. In this sense, mobile phone usage patterns are used to identify and eventually influence human behaviour.

For one, this has the potential to complement traditional statistical surveys. For example, the measurement of poverty levels is one of the politically most relevant statistics throughout the developing world. Using information from mobile phone call records through a plurality of base stations, it has been shown how the prediction of socioeconomic level in a geographic region can be automated (Frias-Martinez and Virseda, 2013; Frias-Martinez and Frias-Martinez, 2014a). Prediction accuracy depends on the kind of variable (for example, predicting gender from mobile phone behavior is surprisingly tricky) (Blumenstock, Gillick and Eagle, 2010; Frias-Martinez, Frias-Martinez and Oliver, 2010). In general accuracy has been shown to be around 80-85% when using mobile call data records like call duration or frequency (Blumenstock, Gillick and Eagle, 2010; Frias-Martinez, Frias-Martinez and Oliver, 2010; Soto et al., 2011; Frias-Martinez and Virseda, 2013). Such ideas can be fine-tuned for cases where more detailed digital footprints are available. For example, given that some 95% of the mobile phones in developing countries are prepaid (ITU, 2014) and given that people put economic priority on recharging the phone, even under economic constraints (Hilbert, 2010), tracking the level of mobile phone recharging provides a powerful proxy source to measure poverty levels in real time on a fine-grained geographic level (Letouzé, 2012).

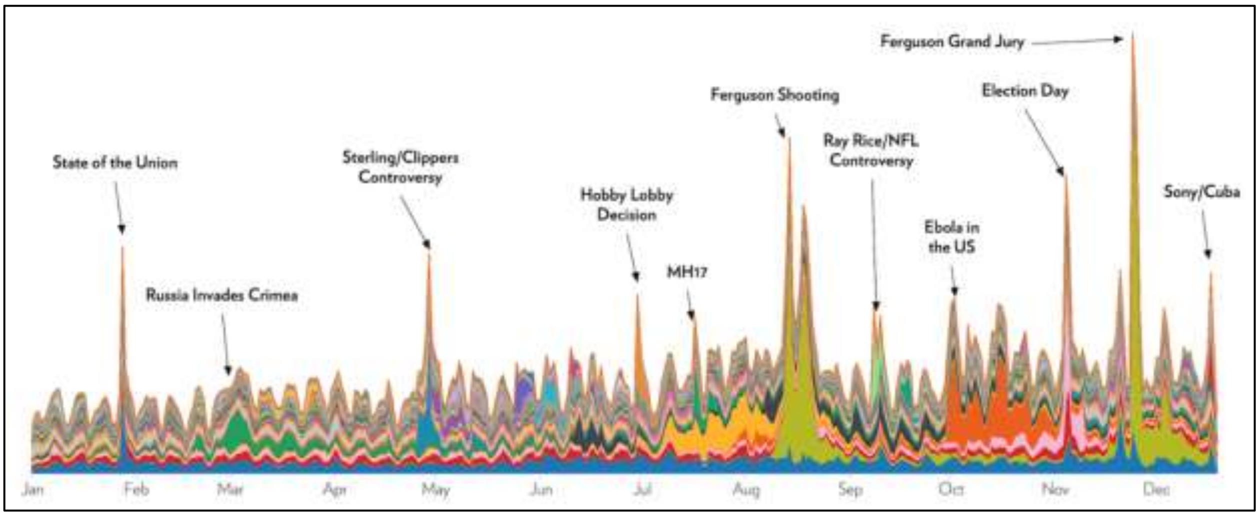
Secondly, universal mobile phone sampling has also demonstrated its potential to obtain new insights. For example, mobile phone records from rural Kenya have been used to provide unprecedentedly detailed travel and migration patterns in low-income settings to understand the spread of malaria (Buckee, et al., 2013) and infectious diseases (Wesolowski, et al., 2014); population movements following an earthquake and cholera outbreak in Haiti (Bengtsson, et al., 2011; Lu, Bengtsson and Holme, 2012); social responses to urban earthquakes in Mexico (Moumni, Frias-Martinez, and Frias-Martinez, 2013); and charity and reciprocal aid among peers in Rwanda after the strike of a natural disaster (Blumenstock, Eagle and Fafchamps, 2012). Telecom companies already sell mobility profiles obtained from mobile phones to business clients, who use it to obtain insights into consumer behavior in real-time (Telefonica, 2012).

B2.2. Real-time shadow. Again, mobile phones play an important role here, as they are not only close to universal, but also provide real-time information. Estonia has been a global leader in the usage of mobile phone location data to complement and to some degree substitute international travel statistics (Ahas, et al., 2011). In Estonia, as in many developing countries, tourism is an important part of the economy. The Bank of Estonia esteem real-time mobile phone traces to give a “reliable overview of persons crossing the Estonian border to travel abroad or to Estonia” and concludes: “the biggest advantage of the method is its speed, laying already existing information stored by mobile operators as potential respondent for statistics. There are neither direct costs associated with the network of interviewers nor a burden for travellers as potential respondents” (Kroon and Pank, 2012). This reduces delivery time to real-time.

But taking advantage of real-time mobile signals is not limited to mobility. The attenuation from radio signals when rain falls between cellular towers has been used as a big data source to measure the amount of rain that is falling in an area (Overeem, Leijnse and Uijlenhoet, 2013). Such real-time and large scale precipitation measurements can bring timely information to farmers, water resource managers, and climate researchers in the developing world where standard rain gauge networks are usually underdeveloped.

Another one of the most common real-time sources for big data is the incessant chatter in online social media. Figure 1 visualizes what U.S. society talked about in 2014, as viewed through 184.5 million Twitter mentions (Echelon Insights, 2014). The ups and downs shown in Figure 1 illustrate that online chatter provides a real-time digital shadow of national interest, outcry, attention, and general sentiment. This real-time source is especially important in developing countries, considering their acceptance and the wide array of content they provide. The top five leading countries in terms of Facebook users in 2013 included India, Brazil, Indonesia and Mexico (Statista, 2014), while in 2011 Kuwait and Brunei had more Twitter users per capita than the UK or U.S., Chile more than Canada, and Brazil more than France or Germany (Mocanu, et al., 2013). Twitter geolocated data has been used for the automatic identification of land uses (Frias-Martinez and Frias-Martinez, 2014b) and the language content of Twitter microblogs has been used to identify the presence of foreign languages and cultural identities, international migration and tourism mobility, including in countries like Malaysia, the Philippines, Venezuela and Indonesia (Mocanu, et al., 2013). Similarly, it has been shown that the 140 character-long micro blogs from Twitter contained important information about the spread of the 2010 Haitian cholera outbreak up to two weeks earlier than official statistics (Chunara, Andrews and Brownstein, 2012). Kalampokis, Tambouris and Tarabanis (2013) investigated 52 articles that made use of social media sources to make social predictions. 13 of them made use of status updates from Facebook and Twitter to predict elections, 10 used web-search engine queries to make economic predictions, and 9 used input from blogs and review boards to predict the spread of disease and reactions to natural disasters. Social chatter becomes a real-time and actionable source of information that can be used to inform decisions in a large variety of contexts.

Figure 1: 184.5 million Twitter mentions in the United States during 2014 as weekly share of news conversation by story, from Echelon Insights (2014)

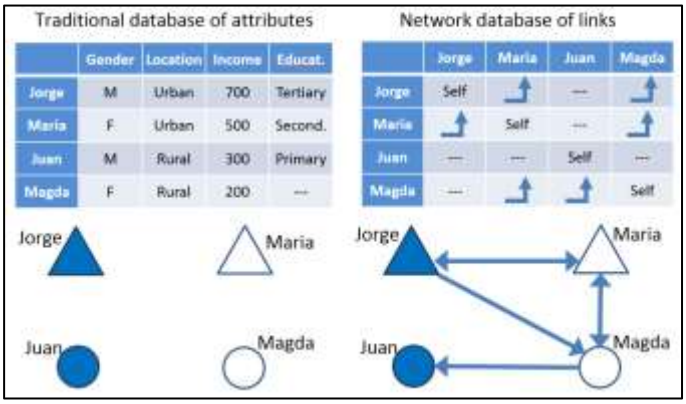


B2.3. Digital footprint as by-product. The production of digital footprints that serve as big data sources is almost inevitable and leads to readily available information sources. Many of the previously mentioned examples bear evidence of the fact that digital footprints are produced as a by-product, without any explicit intention. These cases will not be repeated here. Instead, it is interesting to notice that the digital footprint additionally sheds light on previously under-investigated aspects of the social fabric.

For example, it has provided visible and illustrative evidence of the importance of social ties. Social links have of course always existed: they are the thread that holds the social fabric together. Human civilization has always been one giant network, whereas intricate network structures and its dynamics convert a bunch of individuals into what we call 'society'. However, before the digital footprint, most of those ties were invisible. The recording of such ties involves much more effort than the collection of the demographic and social attributes of an individual, since ties grow exponentially with the number of people in the network. As a result, traditional statistical analysis has mainly focused on collecting and analyzing attributes of independent agents (e.g. income levels, demographics, location, and gadgets that belong to specific social agents). Social network analysis additionally collects and analyses the network ties between these actors. This requires the creation of additional databases that describe social relations in matrix format (see Figure 2). Studying these linkages is complementary to traditional statistical methods. The digital footprint records these exponentially growing statistics anyways, with every digital contact. This provides a wealth of new data about what has been the 'neglected other half' in social analysis over recent centuries: society is not exclusively based on 'who individuals are', but as much on 'with whom individuals are'. The second is often a much more powerful predictor of human behaviour than the first.

The tools and techniques of social network analysis (e.g. Hanneman and Riddle, 2014; Newman, 2010) allow to analyze the networked social fabric that development is made of. Concepts like 'agent of change', 'intermediary', 'gatekeeper', 'broker', 'polarization', 'exclusion', 'marginalization', 'fragmentation', 'discrimination', 'social stability' 'social capital', and 'sphere of influence' are mathematically precisely definable concepts in social network analysis and help to formalize the discussion about development.

Figure 2: Schematization of traditional database (left) and network database (right)



It has been shown that social network ties are important indicators of economic development on the micro- and macro-level. For example, there is a strong relation between economic development and the diversity of individuals' relationships (Eagle, Macy and Claxton, 2010) and of production networks of an economy (Hidalgo, et al., 2007). Network approaches have already proven their usefulness to gain insights and guide policy in ICT4D, such as in the design of media campaigns for reproductive health (Valente and Saba, 1998); the general dynamic of the diffusion of innovations such as ICT (Valente, 1995; Vishwanath and Barnett, 2011); the accelerated diffusion of innovations through pinpointed targeting of opinion leaders (Valente and Davis, 1999), including empirical studies of the acceleration of the diffusion of microfinance (Banerjee, et al., 2012); inter-organizational cooperation and knowledge-sharing within the ICT4D community (Lee and Monge, 2011); and the role of telecenters and cybercafés in the creation of social capital in developing countries (Baron and Gomez, 2012).

The digital footprint does not only make these kinds of social network analyses available, but also makes it actionable. For example, targeting the digital footprint of social connections, the 2012 U.S. Presidential re-election campaign of the incumbent Barak Obama spent some US$ 1 billion with a team of 40 big data engineers (from Twitter, Google, Facebook, Craigslist, stem cell biophysicist, professional poker player) (see e.g. Madrigal, 2012; Woodie, 2013; Rutenberg, 2013). The so-called Project Narwhal created 16 million unique voter profiles, ran some 62,000 computer simulations of likely voter behaviour to rank undecided voters according to a 0 to 10 persuasion score. The campaign paid 35% less per broadcast commercial than the competing Romney campaign (US$90 million less), but rather used the gained insights to present tailor-made campaign promises on social network sites, like Facebook. For example, by posting tailor-made campaign promises for undecided voters on Facebook pages of friends who gave the campaign access to their page (e.g. by being part of the Obama Facebook campaign), social networks can be exploited that involve friendship effects in the message delivery. Insiders claim that the effort changed the voting behaviour of 78% of targeted undecided voters through mass-customized big data interventions on Facebook.

B2.4. Data-fusion. Mobile, social media and network data (and others) can also been combined to fine-tune important indicators of development. For example, the current methods on measuring development are quite coarse-grained, which is exemplified by the most important development barometer, the United Nations Human Development Index (UNDP, 2014). While being a better reflection of development than any single variable, and surely highly innovative 25 years ago (winning Amartya Sen (2000) the Nobel Prize), it merely feeds off four generic indicators: life-expectancy, adult literacy, school enrollment ratio, and gross domestic product per capita. It has increasingly been subject of critique (Stiglitz, Sen and Fitoussi, 2009), and scholars have proposed and developed a large variety of alternatives, including so-called "happiness" indices, which are produced based on costly subject surveys (Frey and Stutzer, 2010).

Big data provides access to a large variety of sources to complement and to even substitute these influential indexes. One important big data source is once again the endless chatter in social networks (see Figure 1). Another one are news items. Using the combination of both sources, the Thomson Reuters MarketPsych Indices (TRMI) distils daily over 4 million social media sites and 3 million news articles through an extensively curated language framework (MarketPsych, 2014). It not only measures different emotions (such as optimism, confusion, urgency etc.), but also opinions (such price forecasts etc.) and specific topics (such as special events, etc.). The company provides 18,864 separate indices, across 119 countries, curated since 1998, and updated on a daily, or even minute basis. The result is a fine-grained, real-time assessment of the local, national or regional sentiment in terms of development relevant indices. Those are grouped into more coarse-grained groups, such as wellbeing, happiness and security, as well as into more fine-grained notions, such as fear, stress, urgency, optimism, trust or anger. These more than 18,000 indices obviously provide a much more fine-grained assessment of the current state of well-being of a society than coarse-grained indices like the Human Development Index and even Happiness Indices, and is able to combine their advantages through data fusion. As mentioned above, the digital big data footprint provides the potential to obtain evidence about socio-economic status, while at the same time providing insights into the emotional state of individuals and social groups.

It is important to mention that the use of diverse digital footprints results in a dataset that is messy and incomplete. In most big data exercises often not one single row of data is complete (not everybody provides social media feeds). However, the explicit data redundancy among different sources in data fusion methods allows to make up for it by the complementary treatment of different sources. As such, data fusion is the assigned remedy for the chronic issue of data incompleteness and messiness in big data.

B2.5. End of theory. Most big data predictions are the result of data mining and machine learning techniques not informed by theory, but simply by pattern detection (Anderson, 2008). The most commonly-used epitome is Google's creation of their natural language translation service without the involvement of a single linguist (Halevy, Norvig and Pereira, 2009). Nowadays Google users can translate among 100 languages. This was achieved by using machine learning algorithms to find patterns in documents previously translated by humans (such as from UN and EU transcripts, and other legal documents). Machines learned patterns that are senseless to humans, but help them to translate language better than previous machines that were trained with theory-laden methods focusing on traditional grammar and semiotics, and even better than many humans. It is the amount of available data that allows such atheoretical machine-learning predictions to work impressively well (Halevy, Norvig and Pereira, 2009).

The amount of socially relevant data also allows machine learning algorithms to become superior to many human judges. For example, it has been shown that “Facebook Likes, can be used to automatically and accurately predict…: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender” (Kosinski, Stillwell and Graepel, 2013). Even more impressive, automated machine algorithms even seem to become superior in classifying the personality of an individual. To test this, subjects were given a standard personality test, and then algorithms were set up to estimate a person’s personality based on Facebook ‘Likes’. Surprisingly “computer predictions based on… Facebook Likes are more accurate (r = 0.56) than those made by the participants’ Facebook friends (r = 0.49); …[and] computer personality judgments have higher external validity when predicting life outcomes such as substance use, political attitudes, and physical health; for some outcomes, they even outperform the self-rated personality scores…” (Youyou, Kosinski and Stillwell, 2015). In other words, machine learning algorithms fed with nothing else than Facebook Likes seem to become superior to the assessments of personality traits and life-outcomes than our friends and even ourselves.

B3. The Ultimate Limit of Big Data for Development

This leaves us with evidence that the atheoretical algorithms of Facebook, Google and Amazon can predict future behaviour better than any psychologist. This is true, but comes with an important, and often neglected condition. It only works if future behaviour follows the same logic as past behaviour. In this case, no theory is required and simple extrapolation based on identified patterns is sufficient. Big data is sufficient. But if the individual falls in love, or changes job or the country of residence, predictions from past data will be limited, if not deceiving. The basic mechanism that produced the pattern changed. However, in that case, a psychologist or an economist who has a theory-driven model will still be able to make predictions, by changing the model’s parameters according to the changed environment.

Changing parameters to explore futures that have not existed in the past requires theory, per definition since those unprecedented future scenarios never existed in empirical reality, but merely exist “in theory”. Such theoretical explorations can be informed by groups of existing data patterns, for example from similar instances. For example, the roll-out of wireless connectivity in a specific rural area might have precedents in different rural areas. However, it never occurred in this area, and each rural area is different. In practice, what is called ‘*ceteris paribus*’ (“all other things being equal or held constant”) never exists. Its assumption can lead to disastrous outcomes, which the well-known result that so-called “best-practices” are often not transferrable. Therefore any adaptation from other cases requires theoretical adjustments, which have to be modeled, be it mentally, through some gut feeling, or formally through formal models. Luckily, the digital revolution also provides formal tools for exploring theoretical parameter spaces and to test for the inclusion of new parameters in a systematic manner.

C. Agent-Based Computer Simulations for Development

Digital technology opened up a new dimension in the endless effort to improve theoretical modeling of real-world dynamics. As early as 1948, Warren Weaver (1948) argued that "science before 1900 was largely concerned with two-variable problems… [while] subsequent to 1900… scientists… developed powerful techniques of probability theory and of statistical mechanics" to analyze a larger number of variables. Weaver concluded that there is a third group of problems that deals "simultaneously with a sizable number of factors which are interrelated into an organic whole… [and] cannot be handled with the statistical techniques so effective in describing average behaviour… These new problems, and the future of the world depends on many of them, requires science to make a third great advance". Weaver already predicted that this new way of doing science will be "effectively instrumented by huge computers". The final breakthrough only came with the digital revolution that enabled computational simulations of these kinds of complex systems some five to six decades later.

The first existing social science computer simulation models were inspired by the work of engineers, who used computers to simulate buildings and bridges that never existed before in empirical reality. Naturally, the main building blocks of those models were the same differential equations and macro-level factors that traditionally rule paper-and-pen-based models (Gilbert & Troitzsch, 2005), such as common in so-called systems theory, including variables like capital, labor, and income. The increase in computational power over the subsequent decades allowed to fine-grain those approaches and saw an increasing transition of simulations "from factors to actors" (Macy and Willer, 2002). Instead of modeling factors of abstract variables, these models simulate individual actors. This allows for the inclusion of much diversity in models, as each actor is a multidimensional variable itself (for example, having specific characteristics in terms of socio-economic, demographic and behavioral characteristics). These more recent models are often referred to as agent-based models (ABMs),also multi-agent models (Epstein and Axtell, 1996; Miller and Page, 2007; Wilenksy and Rand, 2015). Their signature is the possibility to study the emergence of non-linear macro patterns that arise out of a multiplicity of dynamical micro-interactions (Schelling, 2006). The very concept of development, which as a total is more than the sum of its parts, is an example of the kind of social emergence agent-based models aim at replicating.

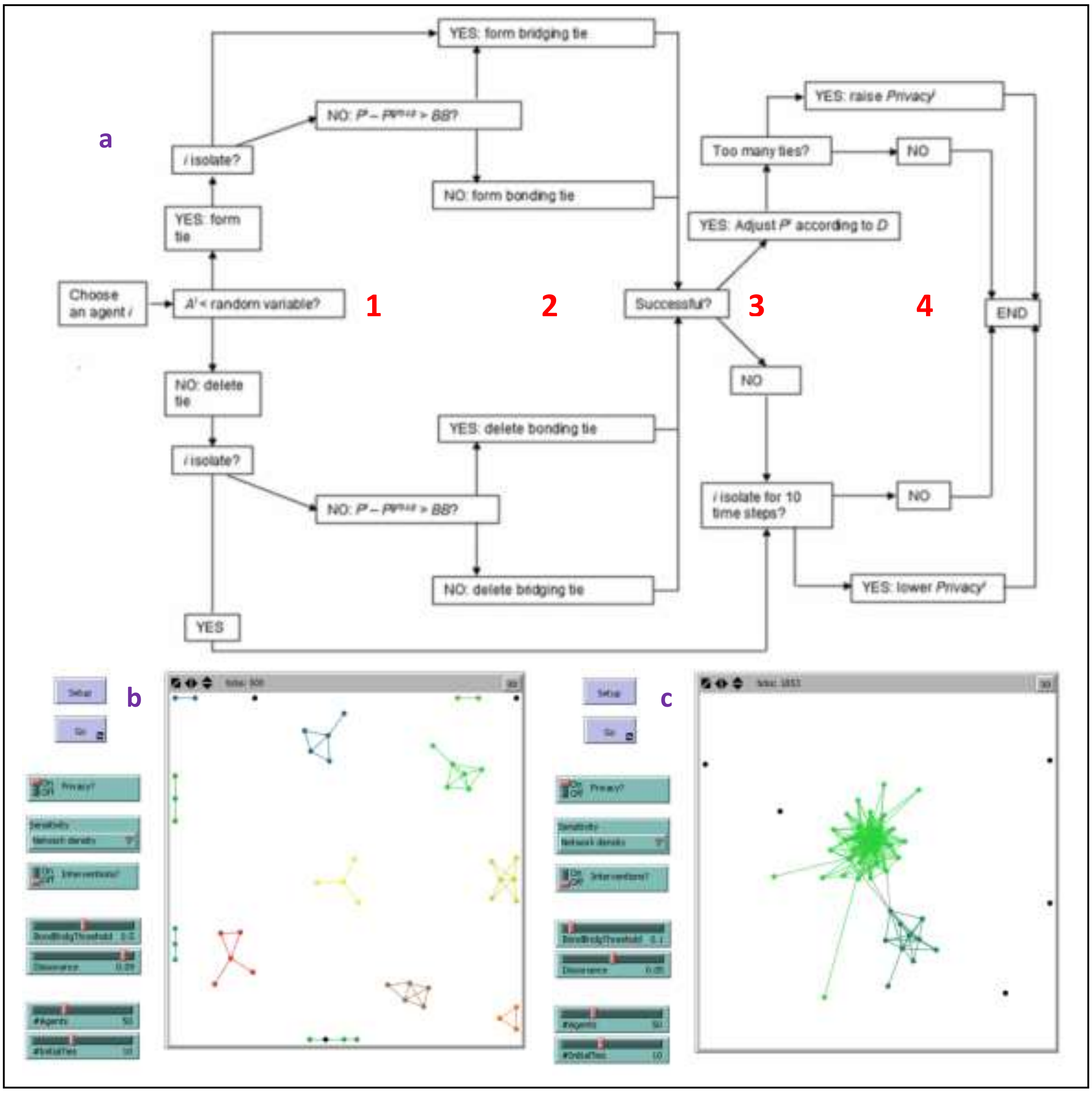
C1. Simulating Future Scenarios

The basic set up of ABMs is reminiscent of popular videogames, such as SimCity, but is executed with scientific rigor. This involves the specification of both agents and their environment, in terms of both attributes and dynamic behaviour. The outcome of the model arises from the interaction effects of these stationary and changing variables. Instead of talking abstractly about them, we start the review by presenting an exemplary use of ABM for the ICT4D-relevant question of privacy in digital networks. Other ICT4D-relevant ABMs exist (for example Lim et al. (2014) built an ABM to simulate the dynamics of opinion formation and fragmentation in a setting of the digital divide). The privacy simulation example is chosen because it illustrates clearly that ABMs allow to build future scenarios for which no empirical data exists.

Tubaro, Casilli and Sarabi (2013) asked "to what extent can we expect the culture of sharing brought about by social media, to drive our societies towards a generalized end-of-privacy scenario—where openness is fully embraced by all as a main norm? And if this is not the only possible scenario, what are the alternatives?" They used the commonly used agent-based modeling software NetLogo (Wilensky, 1999; Wilenksy and Rand, 2015) and created agents with two main characteristics: *privacy*, which quantifies how visible the agent's profile is to others in the network; and *Practices* (P), which quantifies how similar the content or behaviour of the agent is with regard to others in the network. Their environment also had two adaptive characteristics: *Dissonance* (D), which quantifies the general culture of accepting diversity in a society; and a *social capital threshold* (BB), which refers to a general culture of preferring the creation of loose bridging ties, or close bonding social ties that promote trust and support, but at the same time also control and sanctions.

The researchers came up with a sequence of behaviour based on existing theory. The basic process describes individuals that create new social ties based on their similarity with others, and then readjust the visibility of their content based on the extension of their network. Figure 3a presents the proposed behavioral route step by step. While such flow-chart pictures can be daunting at first sight, they are very useful to build the bridge between social behavioral patterns, and the subsequent programming of the respective simulation algorithm. The four main steps (from left to right) can broadly described like this (see Figure 3a): (1) a random coin toss decides if a selected agent builds or deletes a tie. (2) The actual execution of the tie creation/deletion depends on the similarity of the individual and its current group (difference between the Practice (P) values of the agent and its group), which is compared to the social culture of maintaining close or loose ties (the BB parameter). The outcome of this comparison decides about the creation of loose bridging ties, or close bonding ties with others. (3) If a new tie is created successfully, the Privacy value (P) of the individual is adjusted to the attitude toward diversity acceptance (the Dissonance parameter D), accounting for the fact that closely integrated individuals get influenced by common social practice of their surroundings. (4) Last but not least, the agent evaluates its existing ties. In case there are too many ties, the agent lowers its visibility. In case the agent is isolated for a long time, it increases its visibility, lowering its privacy standards.

Figure 3: based on Tubaro, Casilli and Sarabi (2013) (a) Algorithm / behavioral routine for one step in the iterative simulation process. (b) and (c) visualization of simulation results of isolated separate groups and large clustered groups.



This process is iterated for consecutive time steps for many interacting agents. Modeling the result of the resulting interactions might be possible through an intricate set of differential equations in search for analytical solutions. Agent-based computer simulations simply let the algorithms loose and then evaluate the numerical solutions. Figures 3b and 3c visualizes different results of the simulation. The authors find that "interestingly, it is when connectedness is at its highest and content-sharing is most pervasive, that a majority of agents turn their privacy settings on. This remains true even if we assume agents are not excessively sensitive to disclosure and are ready to accept a relatively high degree of exposure of their profiles to others (high thresholds)... Rather, users' reactions are triggered by changes in the structure of their personal network…" (Tubaro, Casilli and Sarabi; 2013). This underlines that privacy "cannot be studied only as the resultant of exogenously given individual attributes such as gender, age or education", but that "the meaning of privacy and its role in our economies and societies are the result of the dynamic interplay of social actors… with their social environment".

The authors went on to study questions like if "the possibility for agents to fine-tune their privacy settings affect the overall configuration of the system?" and if "repeated attempts of online providers to eliminate privacy lead to more or less privacy (through backlashes)?" While a more extensive review is surely beyond the scope of this paper (for more see Tubaro, Casilli and Sarabi; 2013), the point here is to show that empirical data analysis can certainly not answer these questions, since large-scale data about a world without privacy does not exist. Computer simulations can overcome this shortcoming.

While this example is set up for maximal 150 agents and four main parameter variables (see Figures 3b/3c), there is no conceptual constraint to the number of agents and parameters included in the model. Each of millions of individuals can be modeled as a unique case. For example, the spread of infectious disease has been simulated over realistic social networks of California's 39 million inhabitants (including 182 million social network ties) (Barrett et al., 2008). Each individual was characterized with up to 163 demographic variables from census data. 163 variables give a minimal combinatorial space of 10^49 choices, which is of course more than enough to characterize each individual with a unique profile. The computational power and respective algorithms are already developed to scale this up to global dimensions. Back in 2011, simulating a global population of 6.57 billion agents with 2.40 billion infections took less than 8 hours computation time (Parker and Epstein, 2011).

Formally, agents can be defined as computational entities, usually showing some form of bounded rationality (memory loss, nearsightedness, local search), situated in some environment, capable of undertaking flexible autonomous actions with the objective of satisfying their individual or collective need. The involved parameters of the agent and the environment can be fixed or variable (i.e. changing under changing circumstances), which results in a myriad of differential patterns. They are usually derived from a theory and can be calibrated with a combination of previous empirical results. But theory does not only inform the models, models also inform the expansion of theories through the experimentation with parameters (Epstein, 2005). The resulting model-centered science sees theories as family of models, whereas models build a continuous bridge between theoretical hypothesis and empirical testing (McKelvey, 2002).

C2. Characteristics of Agent-Based Models

Summing up, agent-based computer simulation models can be characterized by some five general features. They:

- enable hypothetical *what-if* questions, by exploring theoretical spaces without precedent
- are built on modules of algorithms, which allows for *scalable context dependency* in a cost-effective manner
- are a constant reminder of the *probabilistic nature of all models*.
- *visualize the unbiased reasoning* with which a model carries out the interactive dynamic that arises from model assumptions.
- help *communicate sophisticated and complex findings* to policy-makers and the general audience.

C2.1. 'What-if'?: One important aspect is the flexible testing of policy scenarios. In contrast to statistical (big data) analysis that is fixated on a specific past, ABMs are an excellent tool to answer 'what-if' questions. Simulations can test realities that never existed and policies that were never implemented. This is especially important in the social sciences, since (in contrast to the natural sciences, engineering or psychology), it is not an option to sacrifice or manipulate societies in labs. Artificial societies, however, can be manipulated at will without the approval of ethical review boards.

For example, an ABM of the 2009 H1N1 outbreak in Mexico showed that government mobility restrictions reduced the spread of the virus by about 10% and postponed it by about 2 days (Frias-Martinez, Williamson and Frias-Martinez, 2011). This policy is in agreement with the recommendation of the World Health Organization, which calls for the suspension of activities in educational, government and business units in case of a pandemic. However, such curfews could cost billions. Agent-based models can not only help to quantify the effect of these kinds of policies, but also simulate scenarios with changed variables, such as (even non-linear) effects of the intensity of restrictions (e.g. closing only airports and not schools), or the use of alternative policies. For example, a simulation of the city of Portland has shown that in case of inhalable plague, voluntary mass use of rapidly available antibiotics is as effective as contact tracing, school and city closures (Barrett, Eubank and Smith, 2005).

Another example from the field of ICT4D is the study of the effect of ICT connectivity on collective action (such as in social protests). ABMs can explore hypothetical effects of variables that differ among societies, such as online communication patterns and the distribution of political preferences. An ABM demonstrated that the positive role of ICT in both the level and speed of collective actions is not automatic (as often assumed in so-called "Twitter revolutions" (Morozov, 2009)), but that it greatly depends on and is highly sensitive to the dispersion of participation preferences (Hu, et al., 2014). This suggests that the effect of ICT in collective action is quite different in contexts with dissimilar preference structures among the involved parties.

C2.2. Scalable context dependency: Another advantage of ABMs is their modular flexibility, which provides scalable solutions to focus on concrete problems in specific settings, instead of trying to understand general theoretic tendencies (Banerjee, Banerjee and Duflo, 2011). "This is moving from a general theory which is supposed to be applicable everywhere to very context specific models. Such models can for sure share some common bases but they should also be adapted to the specific context." (Wendelspiess, 2014). Reusing the code allows to create tailor-made models for concrete problems in specific, local- and context-dependent settings. Development landscapes are notoriously heterogeneous, and embracing this diversity can be key to understand subtleties of any intervention.

For example, several ABMs exist to investigate civil violence and riots by simulating the contagious nature of spreading participation and the differential consequences of varying intensities of police presence and reaction time (Epstein, 2002; Goh et al., 2006). Fine-tuned and extended by empirical data, the basic idea behind such models has then been applied to the specific case of the 2011 London riots, which resulted in several deaths and USD 400 million in damages (Davies, et al., 2013). Similar to the creation of different versions of SimCity, the computer simulation of a unique local community can make use of existing software modules, while evaluating a context-dependent future that is different from the past. This provides a cost-effective solution to eventually replace research on 'the representative village in Africa' with 'this specific village in Africa'.

C2.3. Honoring the probabilistic nature of models: Any model is merely an abstraction of reality. Reality is way too complex to be presented in all its detail. A one-to-one model of reality would be a perfectly working model, but is not practical, as it would imply to remodel the universe with the universe. Reducing the number of variables, and coarse-graining over certain details has the unavoidable effect that models become inaccurate. "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful" (Box and Draper, 1987). If different modeler chose different foci in the process of abstraction and reduction, different models will be the result, with potentially different outcomes. It is easy to imagine a slightly different setup for the assumptions of the example-model of privacy above (Figure 3). All theoretical models have this characteristic of reflecting only chosen aspects of reality. As a result, every model is inevitably probabilistic in its nature, as it contains both deterministic rules and probabilistic noise-related deviations from these patterns (Crutchfield, 2012). The noise is the result of a model that does not perfectly specify all initial conditions, and all rules of behaviour of the dynamic. When programming computer simulations, one often does not have any theoretical input on how to program a specific location, behaviour, reaction, or interaction result. As a result, the programmer adds a random choice. For example, in the above agent-based model of privacy, it was randomly decided if an agent would think to add or delete ties. The series of random choices in the model leads to different end-results each time the model runs. They add up to a large-scale 'invariant distribution' of results, but the particular result is the result of path-dependency on initial conditions and the incorporation of 'luck' and 'random choice' in the unfolding of a social dynamic. The explicitness of the probabilistic nature of computer simulations is a humbling reminder that no models can ever make completely accurate predictions.

C2.4. Visualizing unbiased reasoning: Experiments have shown that both laypersons and experts are not bad in setting up mental models and calibrating them in agreement with reality. Any possible bias in this task (e.g. due to selective attention and priming) would also affect the computer-aided modeler. However, formal models are remarkably better than mental projections in carrying out the consequences of the selected and interacting assumptions (Page, 2012). While all models have this characteristics (being one of the main reasons why we use formal models, and do not merely rely on our intuition), the visualizations of computer simulations allow people to track and to appreciate where their intuition leads them astray from logical consequences (often because of fear or the neglect of feedback; Boschetti, 2015).

C2.5. Intuitive science communication: Sticking to the image of SimCity reveals an additional benefit. The multimedia visualization can be used to communicate with, engage, and convince policy makers and stakeholders who lack sophisticated statistical or scientific training. While early simulation software programs like TRANSIMS in the late 1990s were quite sterile (see Figure 4), the application of modern simulations are much more visually rich and can be run on an affordable laptop. In contrast to the intimidating equations and static graphs of traditional analysis, the presentation of dynamic computer simulations is as intuitive as watching and playing a videogame and allows for a rather playful approach to development analytics. Stakeholders can see the social dynamic unfolding and even take ownership of the model by asking for real-time adjustment of parameters in order to test for specific scenarios. Policy-makers can test countless 'what if' scenarios on a concrete setting before taking the plunge for one or the other option.

Figure 4: Evolution of socio-economic simulation software 1999-2013, based on Hyman (n.d.) and Electronic Arts (2014).

D. Opportunities and Challenges for Development

The full potential of e-science becomes clear when combining both the empirical big data and the ABM modeling approach. The challenges become clear when remembering that these innovations are subject to the well-known innovation processes of Schumpeterian creative destruction, including its diffusion and learning curves.

D1. Opportunities

The importance of combining statistical analytics and theory-driven models in the field of development arises from Lucas' (1976) critique. Data from the past have a limited value after an intervention that is purposefully designed to systematically alter the *modus operandi* of the targeted system. It is important to qualify this statement by pointing out that the validity of data insights depends on the kind of statistical analysis. Some tests are designed to be sufficiently broad to predict a large and general group of cases (i.e. testing for 'out-of- group' samples from a more general group of situations), while others are fitted to explain particular cases (Shmueli, 2010). The difference consists in identifying patterns contained in a larger, more general class of cases, and in explaining the particular circumstances of a specific (class of) case. The vast majority of current quantitative social science research focuses on the latter kind of explanatory analysis (mostly executed through significance and R^2 tests) (Shmueli and Koppius, 2011).

However, even statistical tests that test for a very broad family of cases reach their limit when the structure of the system is altered and no data exists for such changed system. In this sense the final goal of e-science is to combine big data approaches with computer-facilitated modeling techniques. "A good complex systems model both begins and ends with data: Low level data is used to formulate the assumptions about the building blocks of the model, and both high and low level data is also used to test whether the resulting emergent phenomena properly correspond to those observed in the real world" (Farmer, 2012).

Often ABMs are calibrated with coarse-grained records (such as done when modeling the process of knowledge diffusion in Santiago de Chile (Piergiuseppe and Taylor, 2004)), or with small-scale survey data, such as done by Wei, Hu and Carley, 2012, who collected some 225 questionnaires to calibrate an ABM that simulates the optimization of m-banking adoption. Simulation models from the natural sciences painstakingly collect data input with sensors and cameras, such as the ambitious Madingley model of the world's ecosystem that aims at simulating "all life on earth" (Purves, et al., 2013). Ecologists report that "the biggest stumbling block… is obtaining the data to parameterize and validate" the model. On the contrary, social scientists do not require "motion-activated cameras… [or] continuous plankton recorders towed beneath ships" to obtain their big data (Purves, et al., 2013). Humans conveniently produce relevant data as a byproduct of their digital life. This allows for a "more realistic representation of human behavior which includes the behavioral changes that might take place" during the dynamic under study (Frias-Martinez, Williamson and Frias-Martinez, 2011).

A quite advanced example is the virtual simulation of the U.S. city of Portland (Barrett, Eubank and Smith, 2005; Hyman, n.d.). This simulation included the modeling of 1.6 million residents with real socio-demographic profiles following identified daily activities in 180,000 specific locations. The data was obtained through a combination of traditional census data, digitally recorded (big data) records and personal activity logs. The results led to surprising insights into complex social dynamics. For example, the city tested for the installation of a new light-rail system (how would traffic patterns and individual behaviour change for different rail routes?). Extensive infrastructure projects like these are also often common in ICT4D.

Summing up in the language of economists, one can say that the approach of e-science "is intermediate between traditional economic theory and econometrics. Traditional economic theory is top-down, modeling decision making from first principles, and then testing against data later… Econometrics, in contrast, takes a bottom up, data-driven, but fundamentally ad hoc approach. The complex systems approach sits in the middle, taking a bottom up data-driven approach that differs from traditional econometrics by explicitly representing agents and institutions and modelling their interactions, without the requirement that everything be derived from fundamental principles" (Farmer, 2012).

D2. Challenges

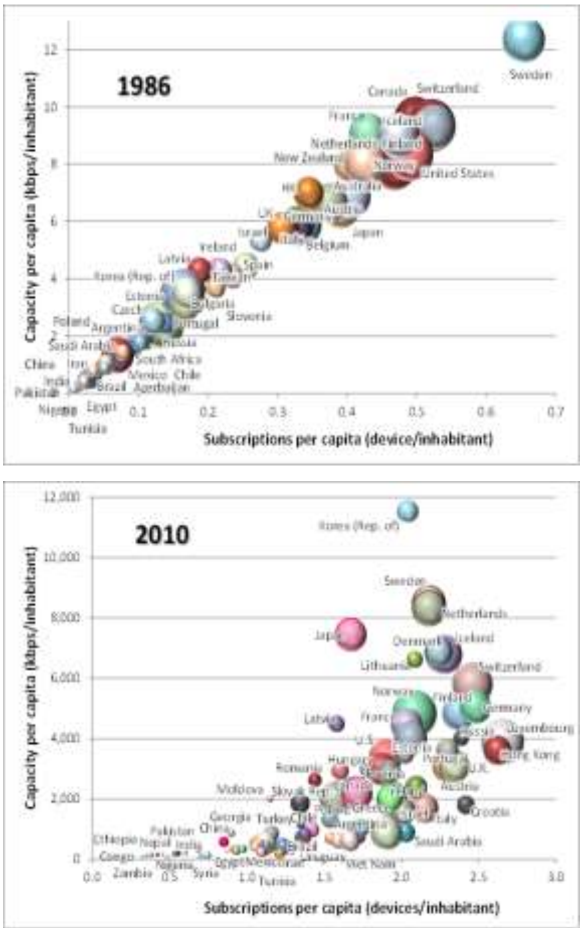
The transition to e-science does not happen in a vacuum, but within the structural, institutional, economic, and social reality of existing societies. This affects the potential benefits stemming from it. The rejection of technological determinism dictates that the application of a new technology will never be inherently and automatically good (Kranzberg, 1986). Emblematic is the case of the digitization of twenty million land records in Bangalore, which created a big data source aimed at benefiting 7 million small farmers in over 27,000 villages (Chawla and Bhatnagar, 2004). Contrary to some expectations, the usual large players were in a much better position to exploit the provided data, resulting in a perpetuation of existing inequalities (Benjamin, et al., 2007). Even well-intended practices can turn out the wrong way. Data is power, and power benefits from data.

Most of the limitations of the use of e-science for development are very similar to the challenges tackled by traditional ICT4D projects, including challenges in the areas of infrastructure, human resources, and institutional frameworks (Hilbert, 2015).

D2.1. Access challenges: First and foremost, the lack of infrastructure access and usage limits the availability of any digital footprint. "Twitter does not represent 'all people', and it is an error to assume 'people' and 'Twitter users' are synonymous" (Boyd and Crawford, 2013). Not surprisingly, it turns out that the question of sample representativeness is closely linked to the degree of digital inequality. Using our previous terms, the question is how close the big data *sampling n* gets to the *universe N*. Blumenstock et al. (2010; 2012) worked with mobile phone data from Rwanda during 2005-2009, when mobile phone penetration was between 2% and 20%. They found that "phones are disproportionately owned and used by the privileged strata of Rwandan society" (Blumenstock, Eagle and Fafchamps, 2012). Frias-Martinez et al. worked with mobile phone big data from a more advanced "emerging economy in from Latin America" (Frias-Martinez and Virseda, 2013) with a mobile phone penetration of around 60-80%. The big data sample matched the social stratification of the available census data impressively well.

However, even once everybody is connected to a mobile phone, the continuous bandwidth divide (Hilbert, 2014) leads to the fact that we will always have better big data sources from some parts of society, but not from others. Over the past two decades, telecom access has ever become more diversified. In the analog age of the late 1980s, the vast majority of telecom subscriptions were fixed-line phones, and all of them had the same performance (see Figure 5). Twenty years later, there is a myriad of different telecom subscriptions with the most diverse range of performances. Far from being closed, the digital divide incessantly evolves through an ever-changing heterogeneous collection of telecom bandwidth capacities (Hilbert, 2014). This provides digital footprints with varying depth through different devices in different countries.

Figure 5: Subscriptions per capita vs. installed bandwidth per capita (in optimally compressed kbps) for 1986 and 2010. Size of the bubbles represents Gross National Income (GNI) per capita (N = 100); based on Hilbert (2014).

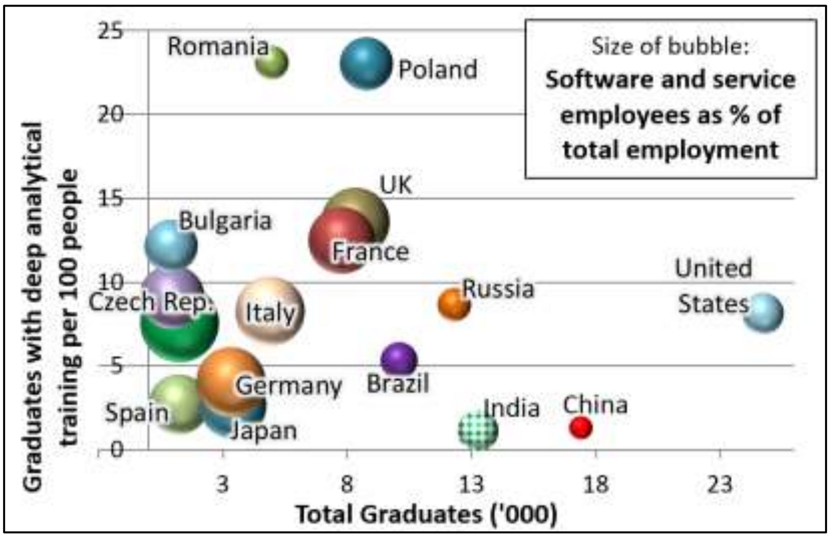


D2.2. Human resource and skills challenges: Hal Varian, chief economist at Google, emblematically stated: “the sexy job in the next 10 years will be statisticians… And I’m not kidding” (Lohr, 2009). The same counts for programmers and computer scientists. Case studies on the use of big data applications in development projects show that adequate training for data specialists and managers is one of the main reasons for failure (Noormohammad, et al.,2010). It is predicted that in the near future 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) (Manyika, et al., 2011).

Within this context of global shortage, from a relative standpoint of international comparison, Figure 6 shows that some emerging economies achieve relatively high graduation rates for professionals with deep analytical skills (high up on the vertical y-axis in Figure 6). In general, countries from the former Soviet bloc (e.g. Romania, Poland, and Bulgaria) produce a high number of analysts. The world’s large developing BRIC countries

(Brazil, Russia, India and China) produce 40% of the global professionals with deep analytical skills, twice as many as the university power-house of the United States (x-axis in Figure 6). They also are expected to have large future demand, as their current share of software and service employees as a percentage of total employment is still relatively low (size of bubbles in Figure 6). Traditional leaders of the global economy, such as Germany and Japan, are comparatively ill-prepared to sustain their demand from national training sources.

Figure 6: 2011 graduates with deep analytical training: total (horizontal x-axis), per 100 people (vertical y-axis); software and service employees as % of total employment (size of bubbles); based on Manyika, et al. (2011); ITU (2014); UNCTAD (2012).



Not only the quantity, but also the quality of analytical training matters. The inventory of big data social media studies by Kalampokis, et al. (2013) revealed that more than one third of the exercises that claimed to demonstrate the predictive power of social media did not even run any explicit predictive analytics (but mere explanatory statistics, such as R^2 analysis). This a better ratio than in traditional (non big data) studies (in traditional information systems research it has been shown that merely 13% of studies that proclaim predictive power use actual predictive analytics (Shmueli and Koppius, 2011)), but still shows systematic misuse of statistical techniques in the social sciences. Given the gradual digitalization of the scientific method, the unavoidable necessity is to assure that young researchers in ICT4D have the required data analytics skills and that more established researchers update their skill set. Collaborations with data-savvy researchers from other fields are another tool to assure that the available big data treasure can still be exploited to ICT4D purposes.

D2.3. Institutional and cultural challenges: Last but not least, the ICT4D community is well aware that any general-purpose technology revolution also requires adjustments in the corresponding institutions and culture: the establishment of what students of technological change call the "new common sense" (Perez, 2004). On the one hand, Schumpeter's creative destruction makes important aspects of previous institutional settings obsolete, while it requires the social construction of new institutions that are up for the digital challenges (Hilbert, 2015; Perez, 2007). Privacy concerns, social discrimination and related abuses are among the biggest threats to the application of techniques like those promoted by big data. A 2014 White House report on big data by the Office of President Obama underlined that big data leads to “vexing issues (big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups)”, while at the same time emphasizing the “tremendous opportunities these technologies offer to improve public services, grow the economy, and improve the health and safety of our communities” (White House, 2014). The challenge to build institutions that minimize the risks and maximize the benefits is especially delicate in developing countries, in which institutional frameworks are notoriously weak, but the catch-up potential is extraordinarily high. On the other hand, the challenge is not merely institutional. It is cultural. It is “not just a question of providing information-, decision-, and action-related skills and other resources for individuals. It will typically require: new, more evidence-based decision-making processes; new, more agile decision-making structures; new institutional values and incentives that orient towards these new decision-making modes” (Heeks, 2014). For students of ICT4D it is not news that technological change requires socio-cultural evolution. This time around it is the socio-cultural evolution of its very own core-business: the application of ICT4D.

E. Fostering the Use of ICT in ICT4D

In the words of fifteen leading scholars in the field: “e-science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. e-science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data… Neither scenario will serve the long-term public interest” (Lazer et al., 2009). The goal has to be to take advantage of the ensuing possibilities for very concrete development projects and policy interventions that benefit the general public. This will not only require the contributions of social scientists, but also the adequate design of information systems that consider and integrate such approaches as an inherent part of their design. For example, information systems could be designed in a way that automatically produces privacy-respecting big data sources that feed computer simulations which enable adjustment of policy interventions on the go.

The field of ICT4D seems very appropriate for such work. On the one hand it counts with a community that is not only convinced of the power of ICT, but also counts with a critical mass of programming and information systems design skills. One in three papers in the field of ICT4D come from computer science and engineering (Hanafizadeh, et al., 2013). On the other hand, ICT4D projects inevitably involve ICT, which will produce big data footprints.

One ongoing challenge consists in significantly increasing the effort to create theory-driven models that can be used in an ever-changing reality. The review in this paper has shown that many more examples and applications are available in the field of big data, as compared to the field of computer simulations. Much more effort has to be put into such theory-driven models. If not, we run the risk of falling into the same traps that Lucas’ colleagues did in the 1970s, some four decades before the big data revolution.

References

Ahas, R., Tiru, M., Saluveer, E., & Demunter, C. (2011). Mobile telephones and mobile positioning data as source for statistics: Estonian experiences. Retrieved from http://www.cros-portal.eu/sites/default/files/S19P4.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

Banerjee, A., Banerjee, A. V., & Duflo, E. (2011). Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty. PublicAffairs.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2012). The Diffusion of Microfinance (Working Paper No. 17743). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17743

Baron, L. F., & Gomez, R. (2012). Social Network Analysis of Public Access Computing: Relationships As a Critical Benefit of Libraries, Telecenters and Cybercafés in Developing Countries. In Proceedings of the 2012 iConference (pp. 377–383). New York, NY, USA: ACM. http://doi.org/10.1145/2132176.2132225

Barrett, C. L., Bisset, K. R., Eubank, S. G., Feng, X., & Marathe, M. V. (2008). EpiSimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In High Performance Computing, Networking, Storage and Analysis, 2008. SC 2008. International Conference for (pp. 1–12). http://doi.org/10.1109/SC.2008.5214892

Barrett, C. L., Eubank, S. G., & Smith, J. P. (2005). If Smallpox Strikes Portland ... Scientific American, 292(3), 54–61. http://doi.org/10.1038/scientificamerican0305-54

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

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & von Schreeb, J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. PLoS Med, 8(8), e1001083. http://doi.org/10.1371/journal.pmed.1001083

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

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

Boschetti, F. (2015). Models and people: An alternative view of the emergent properties of computational models. Complexity, n/a–n/a. http://doi.org/10.1002/cplx.21680

Box, G. E. P., & Draper, N. R. (1987). Empirical model-building and response surfaces. Wiley.

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

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

Campbell, D. T. (1976). Assessing the Impact of Planned Social Change. Occasional Paper Series, #8. Retrieved from http://eric.ed.gov/?id=ED303512

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an Emerging Market. Journal of Forecasting, 32(4), 289–298. http://doi.org/10.1002/for.1252

Chawla, R., & Bhatnagar, S. (2004). Online Delivery of Land Titles to Rural Farmers in Karnataka, India. Presented at the Scaling Up Poverty Reduction: A Global Learning Process and Conference, Shanghai. Retrieved from http://info.worldbank.org/etools/docs/reducingpoverty/case/96/fullcase/India%20Bhoomi%20Full%20Case.pdf

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

Crutchfield, J. P. (2012). Between order and chaos. Nature Physics, 8(1), 17–24. http://doi.org/10.1038/nphys2190

Davies, T. P., Fry, H. M., Wilson, A. G., & Bishop, S. R. (2013). A mathematical model of the London riots and their policing. Scientific Reports, 3. http://doi.org/10.1038/srep01303

Eagle, N., Macy, M., & Claxton, R. (2010). Network Diversity and Economic Development. Science, 328(5981), 1029–1031. http://doi.org/10.1126/science.1186605

Echelon Insights. (2014). #TheYearInNews 2014. Retrieved from http://echeloninsights.tumblr.com/post/105911206078/theyearinnews-2014

Electronic Arts. (n.d.). SimCity Official Website. Retrieved from http://www.simcity.com

Epstein, J. M. (2002). Modeling civil violence: An agent-based computational approach. Proceedings of the National Academy of Sciences, 99(suppl 3), 7243–7250. http://doi.org/10.1073/pnas.092080199

Epstein, J. M. (2005). Remarks on the Foundations of Agent-Based Generative Social Science. Santa Fe Institute Working Papers. Retrieved from www.santafe.edu/media/workingpapers/05-06-024.pdf

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

Farmer, J. D. (2012). Economics needs to treat the economy as a complex system. Berlin: Complexity Research Initiative for Systemic instabilities (CRISIS). Retrieved from http://www.crisis-economics.eu/publication/economics-needs-to-treat-the-economy-as-a-complex-system-2012/

Frey, B. S., & Stutzer, A. (2010). Happiness and Economics: How the Economy and Institutions Affect Human Well-Being. Princeton University Press.

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

Frias-Martinez, V., & Frias-Martinez, E. (2014b). Spectral clustering for sensing urban land use using Twitter activity. Engineering Applications of Artificial Intelligence, 35, 237–245. http://doi.org/10.1016/j.engappai.2014.06.019

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

Frias-Martinez, E., Williamson, G., & Frias-Martinez, V. (2011). An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom) (pp. 57–64). http://doi.org/10.1109/PASSAT/SocialCom.2011.142

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

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics, 55(4), 665–676. http://doi.org/10.1016/j.jmoneco.2008.05.010

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

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

Goh, C. K., Quek, H. Y., Tan, K. C., & Abbass, H. . (2006). Modeling Civil Violence: An Evolutionary Multi-Agent, Game Theoretic Approach. In IEEE Congress on Evolutionary Computation, 2006. CEC 2006 (pp. 1624–1631). http://doi.org/10.1109/CEC.2006.1688503

Goodhart, C. (1976). Problems of Monetary Management: The UK Experience. Hanover New Hampshire: The Public Affairs Center, Dartmouth College. Retrieved from https://www.globalhivmeinfo.org/CapacityBuilding/Occasional%20Papers/08%20Assessing%20the%20Impact%20of%20Planned%20Social%20Change.pdf

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

Hanafizadeh, M. R., Hanafizadeh, P., & Bohlin, E. (2013). Digital Divide and e-Readiness: Trends and Gaps. International Journal of E-Adoption, 5(3), 30–75. http://doi.org/10.4018/ijea.2013070103

Hanneman, R., & Riddle, M. (2014). Introduction to social network methods. Riverside CA: University of California, Riverside. Retrieved from http://www.faculty.ucr.edu/~hanneman/nettext/

Heeks, R. (2014). The Data Revolution Will Fail Without A Praxis Revolution. Retrieved from https://ict4dblog.wordpress.com/2014/08/14/the-data-revolution-will-fail-without-a-praxis-revolution/

Hidalgo, C. A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The Product Space Conditions the Development of Nations. Science, 317(5837), 482–487. http://doi.org/10.1126/science.1144581

Hilbert, M. (2010). When is Cheap, Cheap Enough to Bridge the Digital Divide? Modeling Income Related Structural Challenges of Technology Diffusion in Latin America. World Development, 38(5), 756–770. http://doi.org/10.1016/j.worlddev.2009.11.019

Hilbert, M. (2013). Big Data for Development: From Information - to Knowledge Societies (SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2205145

Hilbert, M. (2014). Technological information inequality as an incessantly moving target: The redistribution of information and communication capacities between 1986 and 2010. Journal of the Association for Information Science and Technology, 65(4), 821–835. http://doi.org/10.1002/asi.23020

Hilbert, M. (2015). Big Data for Development: A Review of Promises and Challenges. Development Policy Review. http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-7679

Hu, H., Cui, W., Lin, J., & Qian, Y. (2014). ICTs, Social Connectivity, and Collective Action: A Cultural-Political Perspective. Journal of Artificial Societies and Social Simulation, 17(2), 7.

Hummon, N. P., & Fararo, T. J. (1995). The emergence of computational sociology. The Journal of Mathematical Sociology, 20(2-3), 79–87. http://doi.org/10.1080/0022250X.1995.9990155

Hyman, M. (n.d.). Part 6 Simulation Models: TRANSISMS and EpiSims. Disaster Resilience leadership Academy. Retrieved from http://www.youtube.com/watch?v=pGftX_56X8g

IEAG (Independent Expert Advisory Group, on a Data Revolution for Sustainable Development). (2014). A World That Counts: Mobilising The Data Revolution for Sustainable Development. United Nations. Retrieved from http://www.undatarevolution.org/report/

ITU (International Telecommunication Union). (2014). World Telecommunication/ICT Indicators Database. Geneva: International Telecommunication Union. Retrieved from http://www.itu.int/ITU-D/ict/statistics/

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

Kolb, J., & Kolb, J. (2013). The Big Data Revolution. CreateSpace Independent Publishing Platform.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802–5805. http://doi.org/10.1073/pnas.1218772110

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

Kroon, J., & Pank, E. (2012). Mobile Positioning as a possible data source for international travel service statistics. In Conference of European Statistics. United Nations Economic Commission for Europe. Retrieved from http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP6.pdf

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

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

Lee, S., & Monge, P. (2011). The Coevolution of Multiplex Communication Networks in Organizational Communities. Journal of Communication, 61(4), 758–779. http://doi.org/10.1111/j.1460-2466.2011.01566.x

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

Lim, D., Lee, H., Zo, H., & Ciganek, A. (2014). Opinion Formation in the Digital Divide. Journal of Artificial Societies and Social Simulation, 17(1), 13.

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

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

Lucas Jr, R. E. (1976). Econometric policy evaluation: A critique. Carnegie-Rochester Conference Series on Public Policy, 1, 19–46. http://doi.org/10.1016/S0167-2231(76)80003-6

Macy, M. W., & Willer, R. (2002). From Factors to Actors: Computational Sociology and Agent-Based Modeling. Annual Review of Sociology, 28(1), 143–166. http://doi.org/10.1146/annurev.soc.28.110601.141117

Madrigal, A. C. (2012, November 16). When the Nerds Go Marching In. Retrieved April 13, 2014, from http://www.theatlantic.com/technology/archive/2012/11/when-the-nerds-go-marching-in/265325/

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

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

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.

McKelvey, B. (2002). Model-Centered Organization Science Epistemology. In J. Baum (Ed.), The Blackwell Companion to Organizations (pp. 752–780). Blackwell Business.

Miller, J. H., & Page, S. E. (2007). Complex adaptive systems. Princeton, New Jersey: Princeton University Press.

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

Morozov, E. (2009, April 7). Moldova's Twitter Revolution. Retrieved from http://neteffect.foreignpolicy.com/posts/2009/04/07/moldovas_twitter_revolution

Moumni, B., Frias-Martinez, V., & Frias-Martinez, E. (2013). Characterizing Social Response to Urban Earthquakes Using Cell-phone Network Data: The 2012 Oaxaca Earthquake. In Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication (pp. 1199–1208). New York, NY, USA: ACM. http://doi.org/10.1145/2494091.2497350

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

Newman, M. (2010). Networks: An Introduction. New York: Oxford University Press, USA.

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. http://doi.org/10.1016/j.ijmedinf.2010.01.002

OECD (Organization for Economic Co-operation and Development). (2013). Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by "big data." In Supporting Investment in Knowledge Capital, Growth and Innovation (pp. 319–356). Organization for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/chapter/9789264193307-12-en

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

Page, S. E. (2012). The Many Model Thinker [MOOC Coursera]. Retrieved from https://class.coursera.org/modelthinking/lecture/115

Parker, J., & Epstein, J. M. (2011). A Distributed Platform for Global-Scale Agent-Based Models of Disease Transmission. ACM Transactions on Modeling and Computer Simulation : A Publication of the Association for Computing Machinery, 22(1), 2. http://doi.org/10.1145/2043635.2043637

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

Perez, C. (2007). Respecialisation and the deployment of the ICT paradigm: An essay on the present challenges of globalisation. In et al. Compano (Ed.), The Future of the Information Society in Europe: Contributions to the Debate. Retrieved from http://www.carlotaperez.org/pubs?s=tf&l=en&a=wprespecialisationandthedeployment

Piergiuseppe, M., & Taylor, R. (2004). Small World Dynamics and The Process of Knowledge Diffusion: The Case of The Metropolitan Area of Greater Santiago De Chile. Journal of Artificial Societies and Social Simulation, 7.

Purves, D., Scharlemann, J. P. W., Harfoot, M., Newbold, T., Tittensor, D. P., Hutton, J., & Emmott, S. (2013). Ecosystems: Time to model all life on Earth. Nature, 493(7432), 295–297. http://doi.org/10.1038/493295a

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

Rutenberg, J. (2013, June 20). The Obama Campaign's Digital Masterminds Cash In. The New York Times. Retrieved from http://www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html

Schelling, T. C. (2006). Micromotives and Macrobehavior (Revised). W. W. Norton & Company.

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

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

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

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

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

Stiglitz, J., Sen, A., & Fitoussi, J.-P. (2009). The Measurement of Economic Performance and Social Progress Revisited. Commission on the Measurement of Economic Performance and Social Progress. Retrieved from http://www.stiglitz-sen-fitoussi.fr/en/documents.htm

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

Tubaro, P., Casilli, A. A., & Sarabi, Y. (2013). Against the Hypothesis of the End of Privacy: An Agent-Based Modelling Approach to Social Media. Springer Science & Business Media.

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

UNDP (United Nations Development Programme). (2014). Human Development Index (HDI). Retrieved September 6, 2014, from http://hdr.undp.org/en/content/human-development-index-hdi

UN Statistical Commission. (2014). Big data and modernization of statistical systems: Report of the Secretary-General (Forty-fifth session). United Nations Economic and Social Council. Retrieved from http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf

Valente, T. W. (1995). Network Models of the Diffusion of Innovations. Hampton Press (NJ).

Valente, T. W., & Davis, R. L. (1999). Accelerating the Diffusion of Innovations Using Opinion Leaders. The ANNALS of the American Academy of Political and Social Science, 566(1), 55–67. http://doi.org/10.1177/000271629956600105

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.

Vishwanath, A., & Barnett, G. A. (2011). The Diffusion of Innovations: A Communication Science Perspective (First printing edition). New York: Peter Lang International Academic Publishers.

Weaver, W. (1948). Science and Complexity. American Scientist, 36, 536–544.

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

Wei, X., Hu, B., & Carley, K. M. (2012). Combination of Empirical Study with Qualitative Simulation for Optimization Problem in Mobile Banking Adoption. Journal of Artificial Societies and Social Simulation, 16(3), 10.

Wendelspiess, F. (2014). Agent-Based Models in Development Economics (SSRN Scholarly Paper No. ID 2397252). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2397252

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

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

Wilensky, U. (1999). NetLogo. The Center for Connected Learning (CCL) and Computer-Based Modeling. Retrieved from https://ccl.northwestern.edu/netlogo/

Wilensky, U., & Rand, W. (2015). An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo. Cambridge, Massachusetts: The MIT Press.

Woodie, A. (2013, June 7). Big Data Analytics Give Electoral Edge. Datanami. Retrieved from http://www.datanami.com/datanami/2013-06-07/big_data_analytics_give_electoral_edge.html

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 201418680. http://doi.org/10.1073/pnas.1418680112