12

BIG DATA AND SOCIAL MEDIA

'Big data' have transformed research practices in the physical sciences, social sciences and humanities. There are now enormous data shadows that reflect a range of important trends, patterns and processes. This chapter outlines some of the possibilities of big data research, with a particular focus on data obtained through crowdsourcing and social media. It then concludes by demonstrating the need for particular caution, but also flagging up opportunities, in the context of development data.

To illustrate the emerging potential offered to development researchers by these advances in technology and datasets, we can look at a UN project exploring the potential links between social media postings and development challenges. The project began in 2011, when a UN report found that messages on the microblogging service Twitter that were discussing the price of rice in Indonesia followed a similar function to official food inflation statistics (UN Global Pulse 2012) (see also http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress). The team began a project in order to broadly understand whether large datasets from social media sites could be used to address populations' vulnerabilities (e.g. affordability and availability of food, fuel, and housing) (see Figure 12.1).

Because it is often extremely difficult to get reliable and timely data about those topics, the researchers turned to data from Twitter. At the time of research, there were over 100 million users publishing 200 million tweets a day. The team wrote some computer code to collect and store data from Twitter that was in Javanese or Bahasa Indonesian. Because those languages are almost exclusively spoken in Indonesia, the research team made the assumption that content in those languages would most probably originate in Indonesia. The dataset was then further filtered to retain only tweets that mentioned buying food (using many synonyms for both 'buying' and 'food').

What the researchers found was a clear relationship between the volume of conversations on Twitter and measurements in food price inflation in official statistics. This could be useful in cases where it is important to understand economic fluctuations quickly. The geography of messages could also provide a useful indicator of where events or processes are happening.

The potentials of this sort of research for development researchers who are often operating in traditionally information-scarce environments are obvious. However, there are also many important caveats and concerns about employing 'big data' in research, all of which are discussed in the remainder of this chapter. Despite these concerns, many

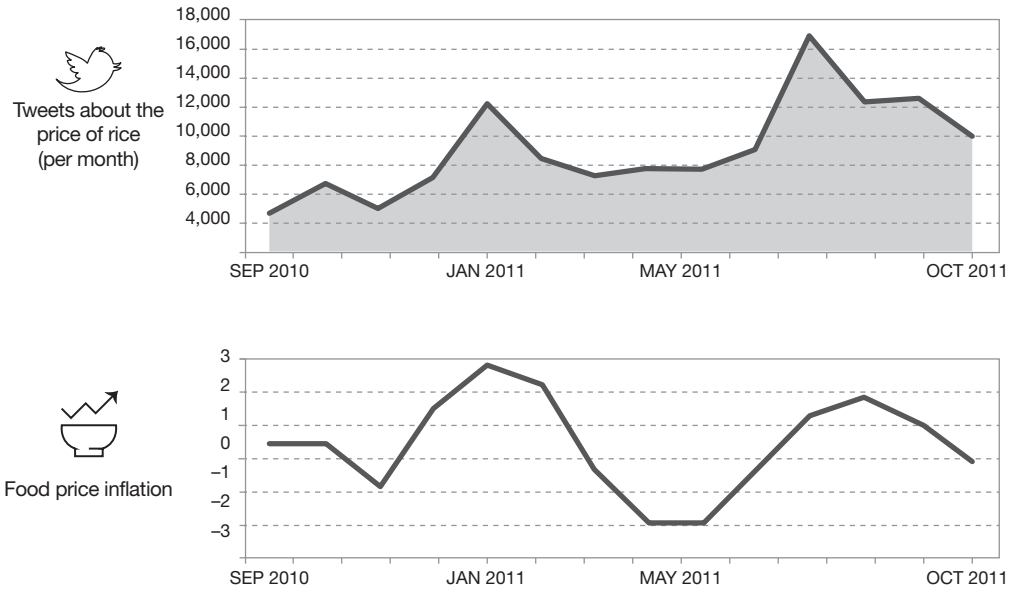

Figure 12.1 Tracking food inflation and tweets about the price of rice

people have argued that 'big data' have transformed the ways in which research is conducted. In many cases, researchers no longer have to settle for a subset or a sample of the data that they are interested in. Instead they can either use the entire population of the trend, pattern or process that they are interested in measuring or use a dataset that serves as a useful proxy for that trend, pattern or process. This chapter guides you through some of the possibilities, potentials and pitfalls of 'big data' for development research. It then offers some starting points for researchers interested in doing big data: reviewing some of the key tools and strategies that can be employed by novice big data researchers. To give you a sense of the scale and range of issues that you need to consider in developing big data-based development research and an inkling of the range of topics you could address using this approach, Charles Brigham (Box 12.1) reflects on his experience of using crowdsourced and big data in his research.

First, however, it might be useful to clarify what it is we mean by 'big data' and social media. The term 'big data' does not have any agreed definition, but generally is used to refer to datasets that have a *volume*, *variety*, *velocity* and *veracity* at a different scale from that researchers are used to dealing with.

The *volume* of many datasets has reached a point that they are now so large that they are difficult to handle with traditional research tools (such as Excel). The *variety* of data refers to the fact that data can take a range of forms (e.g. structured, unstructured, text, non-textual data). 'Big data' can also refer to both directly captured data (i.e. data that exist because the goal behind their creation was to perform some sort of direct measurement) and 'exhaust data' (data that are produced as a by-product of a process or system rather than being the main purpose of a process or system, frequently also referred to as 'trace data'). *Velocity* indicates that many data are now captured in motion. That is,

Box 12.1 Charles Brigham: Using crowdsourced data and big data in development research

Outsourcing tasks to a distributed undefined public, rather than a specific body, is becoming more commonplace in development work. Processes of sourcing data through crowdsourcing have evolved beyond mere collection of data at a generic level involving people at large, to becoming a strategy for gathering collective development intelligence to complete specific tasks. This, coupled with 'big data' systems, enables the capture and processing of massive amounts of near real-time information, making it accessible and providing an opportunity to significantly change the way we create, execute and communicate development. It indicates a movement away from an information age, with lots of data, to the idea of a knowledge age, where information and data are converted into knowledge through a collective crowdsourced voice. Consuming information and collaborating to become educated brings us closer towards the coming age of correlation of information, in parallel with stronger visualization and storytelling.

Big data tools, such as Hadoop (http://hadoop.apache.org/), process data quickly and make it useful, allowing us to ask questions and get the answer back before we've forgotten why we've asked the question in the first place. This is what is fundamentally new in using crowdsourced and big data in development and is only recently becoming widely understood. Past development work has focused on utilizing personal, siloed computing for timely processing of data, information and common social data channels to filter out noise and understand crowd voice in development projects, often resulting in insufficient gain. Today innovations in cloud computing, SAS and data science are providing the backbone for tools to allow big data for better informed development decisions. A key challenge will be in realizing potential and making these trained systems easier to build and maintain.

Business and marketing sectors are utilizing big data and crowd voice to tackle technological complexity, data accuracy and rights of use, business and technological alignment and need for data specialists. It is hoped that the development field will follow suit to strengthen collective intelligence, crowd voice and big data to utilize this wisdom within development institutions. We see this emerging through initiatives such as open data, greater access to information and improved technology tools and uptake but there is still much further to go in order to utilize new methods in order to impact the poor and marginalized. As an example, Indonesia's National Program for Community Empowerment (PNPM), is the largest community-driven development project globally. It has continually utilized social tools (networks and platforms) to enhance collaboration. While a massive number of users are linked to these networks, they in no way constitute or foster an active crowdsourcing for development environment: less than one-quarter of the thousands of locally driven social networks and groups at the community level are annually active. Over recent years fewer people are utilizing these common social networks to collaborate. *This is being addressed in real time.*

Newer beneficiary engagement platforms are emerging to utilize crowd voice and big data. Currently the Indonesian community-driven development project mentioned above is utilizing a multi-mode approach to collecting feedback from citizens (i.e. mapping, SMS, mobile and web) for implementing and development partners around project-specific development outcomes. The aim of the platform is to improve the current mechanisms of citizen input and feedback loops. The streamlining and regular sharing of large volumes of data at small area geographies will effectively engage citizens to provide actionable feedback and provide tools for multiple stakeholders (project implementing units, government, service providers and relevant staff) to view, act on, monitor and analyze crowdsourced and big data as needed.

data are not in neatly stored databases, but rather captured as traces from ongoing events and processes. *Veracity* is a way of describing the uncertainty and complexity in much 'big data' because of imprecise and unreliable data and data types, uncertainty in unstructured data from conversions and probabilities used in machine learning and uncertainty in how good a proxy the 'big data' that we have are for the more general phenomena that we wish to measure.

The reason that many of these data exist is because ever more facets of everyday life are both digital and digitized. Those everyday activities, transactions and behaviours together are encoded into datasets of sizes that most researchers have never had to deal with before. 'Big data' are also not limited to activities in rich countries. The high penetration of mobile phones around the world has meant that billions of people around the world have turned into digital sensors: emitting and sharing data. Telephone metadata can reveal movements and socio-spatial interactions (e.g. the number of people in one town who communicate with people in another town).

Many of the 'big data' available to researchers come from social media platforms. Social media can be particularly useful because they offer a platform for people to share information about their lives on. This information, in the aggregate, can potentially offer social, economic and political insights that would otherwise be challenging to gain. 'Big data' are not only volunteered, but are sometimes unknowingly emitted. People riding transit systems, applying for permits, paying tolls, triggering sensors, agreeing to terms of service (without reading those terms) and performing all sorts of other acts can all contribute to repositories of 'big data'. The question is what can we do with those data and how can they be useful to development researchers?

Why use 'big data'

The overarching reason that we might want to use 'big data' is because of an ability to look at broad-scale development-related patterns and processes that either are not captured by traditional sources (such as the World Bank) or are time sensitive. It is often the case

that datasets of interest to development researchers do not exist, are of poor quality or are rendered inaccessible by governments or companies. Other times, high-quality and accessible data exist and are available, but are hopelessly out of date because of quickly changing contexts. A growing number of researchers are therefore looking for alternate types of direct and exhaust data that allow them to still ask broad-scale questions about patterns and processes related to development.

Much of the power of 'big data' also comes from the ability to combine a diverse range of sources in order to look for connections and associations. Those connections, in turn, can reveal patterns, processes, and relationships – that might otherwise remain undiscovered. Said differently, much 'big data' work is about exploration and obtaining insights that might be explored in more depth with other methods (albeit within the framework of an informed understanding of the field of study to be able to formulate good hypotheses and questions).

Some examples of contexts in which 'big data' might be used in the context of development research include:

Public health – an analysis of mobile phone traces (and thus, the flows of human mobility) could reveal how diseases and illnesses spread.

Transport planning – measuring flows of minibuses to understand the supply and demand of public transport.

Disaster response – user-generated reports could be used to create almost real-time understandings of human needs in information-scarce environments.

Demographic data – mobile phone usage combined with GNI or other metrics of human development could help to create up-to-date understandings of population in non-census years.

Understanding society – by knowing who communicates with whom, researchers can gain detailed insights into the micro-contours of different social groups: thereby understanding who occupies central and who occupies marginal positions.

Preventing famine – monitoring sudden price spikes in markets may lead to identification of early warning signs. In future cases, these early warning signs could prompt intervention before prices raised too far.

Economic justice – databases of market prices can also be used to allow farmers and other producers of goods to better understand the worth of their products and therefore avoid exploitation by intermediaries.

Many other examples exist, but already you can see that a core value of 'big data', especially in the context of development, can be illuminating needs and patterns in the context of previously information-poor environments.

How to use 'big data'

How then do you begin using and analyzing 'big data'? Below we offer a few strategies and tools. Some of this guidance is directed towards the novice, while other suggestions will be more useful to researchers with at least some experience writing computer code with a computer programming language. We address where you might obtain data, how

Box 12.2 Emmanuel Letouzé: Big data for development: how can we make sure big data are not the 'new oil'?

Over the past few years the concomitance of the availability of passively emitted data about human actions, generated and collected digitally as a by-product of the analogue-to-digital transition and of advances in computing and analytics capacities – dubbed 'big data', has given rise to questions about the phenomenon's potential to foster human progress. Publications, discussions and initiatives around 'big data (or data science) for development' (or 'social good') have flourished, culminating in the recent call for and numerous references to a 'data revolution' (with or without the qualifier 'big' in front). By and large, this call reflects and nurtures a salient recognition of the need for more agile and accountable policies and a widespread dissatisfaction with current systems and tools in the age of digital data and technology. So for some, big data are thought to be set to become the 'new oil' that provides, for human development, an immense source of value if only we are able of to refine it. There is some well-founded scepticism about the hype around some of these initiatives. If anything, the fact that the 'old oil' hasn't done much to foster human development among the population sitting on it (which has been coined the 'resource curse') may serve as a cautionary tale of some of the risks involved.

But what exactly are big data and what are these risks? In recent papers and interventions, I first distinguished big data and 'Big Data', referring to the former as data and to the latter as a field. Then, I argued big data's novelty is not 'about' size but about the availability of what we (with co-authors Patrick Meier and Patrick Vinck) have described as 'digital traces of human actions picked up by digital devices' (2013). Professor Sandy Pentland (2012) focuses on what he calls 'digital breadcrumbs' or data particles that individuals leave behind. The main message is that big data as data are essentially a *qualitative* shift. Subsequently, citing Gary King, I contend that *big data are not about the data*, but about the analytics, which I further interpret, following Kentaro Toyama, as being about *intent and capacities*. As such, big data are or should be interpreted and approached as a deeply political phenomenon.

Big data's potential to enhance 'our' understanding of and ability to predict mobility patterns, poverty dynamics, and more, is evident, as documented in a fast growing body of research. There is a detrimental dearth of data in many developing countries and big data may help fill part of a gap which, in the case of Africa, has been referred to as a 'statistical tragedy'. But the downside, and the main risk, is the belief that lack of data is the main issue impending human progress and the temptation to 'leapfrog' into a data-driven model of governance designed by and for an enlightened minority. The corollary risks are the creation of a new digital divide, dehumanization or de-democratization of decision making, and/or infringement on fundamental human rights – with respect to privacy and confidentiality most evidently. Three main sets of related factors create or increase these risks. One is a normative and empirical technological bias with insufficient consideration

for the phenomenon's humanistic (including ethical) implications. Two is poor institutional connectivity between traditional policy actors, on the one hand, and data, web and computer scientists and ethicists, on the other hand. Three is limited structural channels and technical capacities for local actors to be fully engaged and access these resources.

So the fundamental question my work is trying to tackle is: what would a humanistic big data revolution – one informed by the principles and attempting to contribute to the objectives of human development – look like and take? How could I better identify and contribute to avoiding the traps that may impede its realization? This is a huge but exciting undertaking.

you might analyze them and how you might gain relevant skills or assistance. Indeed, as Emmanuel Letouzé outlines in his reflection in Box 12.2, the issue of analytics is vital in approaching work with big data and that care is needed to ensure the field is developed in a humanistic way – in others words, one that is both ethical and contributes to human development.

Where do your data come from?

'Big data' could be thought of as one (or more) of three varieties: machine, social and transactional data. Machine data include data gathered from machines or sensors. It could also refer to a variety of exhaust data such as web logs or mobile phone traces. Social data can refer to mining the traces that people leave behind on large social media platforms (e.g. Facebook likes or Twitter mentions) or data that they actively send and contribute (such as emails, free-form text, images, audio, videos). Transactional data include records of things like logs of processes, emails, stores of documents and a range of other transactions. Following the link at http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_ Briefing_2012.pdf provides another way of thinking about the three primary types of 'big data' source: individual (e.g. data exhausts and crowdsourced information), public (e.g. tax, facility or census data) and private (e.g. transaction data or spending data). These data will tend to be in one of three forms: *Precompiled datasets*, *structured data*, *unstructured data*:

Precompiled datasets: Measurements are already defined and data have already been collected and compiled by other researchers or organizations.

Structured data: Here there tends to be a need to compile the dataset, but the measure of interest is usually already defined. Often accessible through an application programming interface (API).

Unstructured data: Here raw data are gathered via an API or other sources, but then must be transformed to create the measure of interest (e.g. primary language of a speaker or tweets about price of rice). In other words, a proxy often has to be defined in such cases.

In some cases, you may be able to obtain such data simply by asking. Many government agencies, NGOs or firms might welcome the analytical expertise that a researcher could bring to their organization. However, because 'big data' often tend to contain sensitive, private and personal information, organizations can be reluctant to allow access to their information. If you do gain access to 'big data' in this way, you will usually be asked to sign some type of data-sharing agreement. Here it will be important to ensure that any agreement you sign provides you with the intellectual freedom to ask and answer your core research questions, and publish and share your results. There is no point in spending time carrying out research if it is only going to be embargoed by an organization that does not like your findings.

When data are not easily available from the organization that collects them, you may have to turn to programming or scripting to obtain the information that you need. Many repositories of data will have an application programming interface (API), which is a set of procedures, instructions and standards used by computer programs to request services and data. In other words, it is a set of building blocks that you can use to customize the way in which you extract the specific types, forms and quantities of data that you need (meaning that you do not have to reinvent the software wheel for common uses).

Many of us would not think twice about learning a local language before doing fieldwork and yet relatively few development researchers in the social sciences and humanities know how to 'speak' a scripting language. (Scripting in this context refers to the short computer programs that can be written to perform relatively narrowly defined tasks.) Luckily, learning the basics of a scripting language (such as Python, Perl, Javascript or Ruby) is now a relatively accessible task. Universities and colleges around the world offer short courses that allow programming basics to be understood in a few weeks. Many such courses are designed for people who have full-time jobs and so tend to either be evening and weekend courses or short, intensive programmes. Other learning resources include:

codeacademy.com – an interactive platform that offers free courses on a variety of programming and markup languages. The site is highly popular is has a very broad user base (five million users).

Coursera.org – another platform offering introductory programming courses designed in/by universities such as Stanford and Toronto. The courses themselves are free, but the organization charges fees for certificates to prove that you have done the work.

Khan Academy (khanacademy.org/cs) – features a series of popular video-based programming 'micro-lessons' in a range of languages.

Open Courseware Consortium (ocwconsortium.org/courses) – also offers dozens of free courses on programming. These full-length courses are usually designed by large educational institutions such as MIT and the University of Cape Town.

Finally, we want to mention two alternate ways of collecting data, both of which involve outsourcing tasks to distributed workers. First, there are now a variety of online job marketplaces (such as odesk.com and elance.com) that mediate interactions between people who need customized code and people who can design customized code. Although

using such platforms can often be a quick and cheap way of obtaining a customized software tool to gather data, it is also important to recognize their limitations. The most fundamental one being that it is important to always understand how your datasets are put together. If you are unable to decipher the code that you are using then it is quite possible that you won't ultimately understand many of the nuances of your own data.

Distributed workers could also potentially be employed in a fundamentally different way as 'paid crowdsourcers', 'microtaskers' or 'clickworkers.' These terms imply that workers (hired through sites such as Amazon's 'Mechanical Turk' or crowdflower.com) can be paid to perform large numbers of small tasks. For instance, those paid crowdsourcers could be paid to manually put together databases of all health clinics in East Africa or they could be paid to standardize addresses in a large database into a format that allows them to be mapped. Here, again, it is important to proceed with caution. Many microworkers are based in the global south and are often paid exploitatively low wages for their labour. While microwork platforms allow, and perhaps even encourage, cheaply sourced work, researchers should make sure that they have considered what a fair salary might be.

Where does your analysis power come from?

'Big data' analytics can take a variety of forms depending on the specific data, frameworks, and questions that you have. Querying, mining, visualizing, modelling, simulating, natural language processing, geospatial analysis and voice/video analytics are some of the most common tasks that 'big data' are enrolled into. Many (but not all) of these methods employed in 'big data' analysis are existing methods that are applied at a larger scale. Below we describe some of the tools, infrastructures and methods useful to 'big data' research.

Computing cluster

In some cases, the speed and size of your own computer may be sufficient to process the data that you want to analyze. However, for many other 'big data' tasks, you need to turn to more powerful resources. Computer clusters are groups of machines that work together to solve tasks that involve large amounts of processing power or disk space. Open-source software frameworks such as *Hadoop*, which can manage tasks in clusters, have made it possible to analyze massive datasets without the availability of expensive computing resources. Solutions like Hadoop allow large datasets to be processed in a distributed manner: meaning that you could set up a 'Hadoop cluster' on a handful of computers (or, of course, many more). None of those computers needs to be high end or particularly powerful, but their combined resources (linked through the Hadoop framework) means that powerful computing resources can be made available for processing huge datasets.

Hosted cloud computing

Hosted cloud computing can be employed in a similar way to the computing clusters described above, the main difference being that you do not have to build the system yourself (however, this is not to say that the setting-up process is always straightforward). A variety of companies offer researchers the ability to outsource such tasks and pay (usually relatively small amounts) for the resources used. Amazon Web Services is one of the most used services in this category (Google's App Engine, and Microsoft's Azure offer similar services).

Cloud labour

A very different way to analyze large datasets is to rely on what has been dubbed 'cloud labour'. The term refers to a globally distributed labour pool of people willing to do small defined tasks for very small payments. This means that you could take a very large task that would be hard for a computer to do automatically (for instance, classifying tens of thousands of images) and break the task up for people around the world to help you with. Crowdflower and Amazon's Mechanical Turk are two of the largest marketplaces for cloud labour.

Positionality and ethics

Much has been said about the transformative power of big data in the contexts of development. However, it is also important to recognize some of the potential problems and pitfalls of 'big data'.

As the size, speed and scale of your data increase, your ability to adequately contextualize data often decreases. This can lead to spurious correlations and misinterpretations of results. Not only can the signals in our data be noisy (and false), but those signals could serve to misdirect us from more fruitful lines of inquiry. The saying that 'if your tool is a hammer, then every problem looks like a nail' is especially apt in the context of 'big data'-driven research. Many of the most important questions in development might not be able to be answered by large datasets and instead will require in-depth, contextual knowledge. In other words, 'big data' themselves can rarely substitute for other rigorous quantitative and qualitative research methods, but rather can be a useful complement.

'Big data' also raise a range of important privacy and ethical issues. Data are sometimes released in (or about) low-income countries that would not be released in (or about) rich countries. The EU, for instance, has relatively strict laws about the ownership, use and dissemination of personal data. But such laws do not exist in most of the rest of the world. It is therefore important to reflect on not just whether data are available but whether you actually want the data that are available (they may for instance contain sensitive information), or whether their use could be considered ethical (see Chapter 5 for a more thorough treatment of ethics in research). The Association of Internet Researchers has an

up-to-date and evolving ethics guide (http://aoir.org/documents/ethics-guide/) that keeps track of developments related to technology and ethics – that could serve as a useful starting point for questions about 'big data' and ethics.

Finally, there are also concerns, alluded to earlier, about working with distributed microworkers. Many of these workers are based in Bangladesh, the Philippines, India and other low-income countries and work under poor conditions. It is your role and duty to ensure that you are constantly carefully reflecting on the wider impacts of your footprint in the world as a researcher: a task that becomes more difficult when interacting with people through relatively opaque internet platforms.

What to take from this chapter

'Big data' could serve as a way of measuring the pulse of society, tracking trends that were previously hard to measure and categorising and analyzing messy and large datasets. Some have gone as far as to claim that 'big data' represent the end of theory; that 'big data' can potentially generate more accurate or true results than specialists who traditionally craft carefully targeted hypotheses and research strategies. However, caution is essential. Understanding the social, economic, and political contexts under which data are produced will always matter (Graham and Shelton 2013). 'Big data' will rarely be able to offer a pure and unbiased reflection of any pattern, place or process, either because so much 'big data' comes in the form of closed, restricted or black-boxed datasets or because the 'big data' being used are not actually an appropriate proxy of the pattern, place or process of interest.

As Escobar (2008) has noted, the very processes of measuring and having data has in many ways created practices of development. As ever more facets of development are quantified and datafied, it will then remain crucial to consider not just what 'big data' measure and what questions they allow us to answer, but also what they omit and how to keep a focus on some of the most crucial and critical questions in development.

In sum, the ready availability of massive datasets and the tools and methods to analyze them is beginning to have (and will undoubtedly continue to have) transformative effects on research in the contexts of development. This chapter has offered a starting point for researchers who are interested in reflecting on how 'big data' might be best embedded into their own work.

If you want to explore the ideas covered in this chapter in more detail, you may find these further readings useful:

Escobar, A. (2008) 'The problematization of poverty', in Chari, S. and Corbridge, S. (eds.) *The Development Reader*. Oxford: Routledge.

Graham, M., Hale, S. and Gaffney, D. (in press) 'Where in the world are you? Geolocation and language identification in Twitter', *Professional Geographer*.

Graham, M. and Shelton, T. (2013) 'Geography and the future of big data; big data and the future of geography', *Dialogues in Human Geography*, 3: 255–61.

Hilbert, M. (2013) 'Big data for development: from information to knowledge societies'. SSRN Scholarly Paper No. ID 2205145. Rochester, NY: Social Science Research Network; http://papers.ssrn.com/abstract=2205145.

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

UN Global Pulse (2012) 'Big data for development: challenges and opportunities'. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-GlobalPulseMay2012.pdf.

World Economic Forum (2012) 'Big data, big impact: new possibilities for international development'. http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development.