GeoJournal (2015) 80:503–518
DOI 10.1007/s10708-014-9603-5

Is bigger better? The emergence of big data as a tool for international development policy

Linnet Taylor · Ralph Schroeder

Published online: 11 October 2014

Abstract The use of digital communication technologies, and of mobile phones in particular, has seen an exponential rise in low- and middle-income countries over the last decade. These data, emitted as a byproduct of technologies such as mobile phone location information and calling metadata, have the potential to fill some of the problematic gaps in data resources available to country policymakers and international development organisations. Using three examples of current big data initiatives in the international development field, we examine the implications of these new types of data for development policy and planning: their advantages and drawbacks, emerging practices relating to their use, and how they potentially influence ideas and policies of development. We also assess the politics of these new types of digital data, which are often collected and processed by corporations or by researchers in industrialised countries. Our analysis indicates that these new data sources already represent an important complement to country-level statistics, but that there are currently important challenges which will need to addressed if the promises of big data in development are to be fulfilled.

Keywords Big data · Development · Policy

Introduction: ‘missing’ data

There has been an exponential increase in the use of digital communication technologies in low- and middle-income countries (LMICs[1]), from mobile phone usage to mobile and fixed broadband internet (ITU 2013a). To take the Sub-Saharan African region as an example, the number of mobile phone subscribers has increased at a rate of 18 per cent annually since 2007, reaching a total of 253 million in 2013 (GSMA 2013). The development of content and applications targeted to people in LMICs is increasing the relevance of these technologies, and the consequent traces and signals emitted through people's use of them have the potential to make LMICs' citizens a rich mine of information about health interventions, human mobility, conflict and violence, technology adoption, communication dynamics and economic behaviour (for an overview,

L. Taylor (✉)
University of Amsterdam, Plantage Muidergracht 14,
1018TV Amsterdam, The Netherlands
e-mail: l.e.m.taylor@uva.nl

R. Schroeder
Oxford Internet Institute, 1 St Giles, Oxford OX1 3JS, UK
e-mail: ralph.schroeder@oii.ox.ac.uk

[1] We use the World Bank's definitions grouping countries, see: http://data.worldbank.org/about/country-classifications, where LMICs have incomes of $1,036–$12,616 and high income countries (HICS) above that threshhold. Our particular focus is the low- and lower-middle-income countries, with an upper threshhold of $4,085 per capita, which includes India and most of Africa.

see Lane et al. 2010; Netmob 2013; Blumenstock 2012; Bengtsson et al. 2011). These digital traces are becoming seen by policymakers and researchers as a potential solution to the lack of reliable statistical data on lower-income countries (Mann 2013).

Given the significant cost and local capacity-building necessary to improve country-level data collection (Jerven 2013), using large-scale, born-digital datasets may provide proxies for some indicators important to international development policymakers, making it an increasingly attractive option. [We use the term 'born digital data' as subset of big data to denote data that are digital from the start rather than starting out in non-digital form (see Borgman 2014)]. Access to these new forms of data is thus an ongoing priority for local and international policymakers alike, as well as for researchers interested in understanding trends in economic and social activity in these countries. This paper will explore the increasing use of big data in the field of international development to inform policy and interventions. The terminology of 'big data' has also become highly popular in the business press and among policymakers: here, we can define the term pragmatically as allowing a step change in the scale and scope of what can be known in relation to a given phenomenon or object (Schroeder 2014). We will illustrate its use through recent examples of data science in the development and humanitarian sphere, with particular attention to the issues they raise to do with accuracy, validity and contextualising these types of data.

Data about social patterns in LMICs and elsewhere have often been scarce. Jerven, in his book on African development statistics (2013), has shown that not only are reliable data missing on Africa in particular, but just as importantly, the resources to gather those data are generally absent in LMICs, and are unlikely to emerge in the near future due to the cost and capacity-building challenges involved. Jerven calls for much more research, and also qualitative and interdisciplinary research, to improve the data, arguing that it is critical for populations in places like Africa to be counted, and that some of the data subjects are aware of this: to be counted means potentially obtaining access to resources. This potential is particularly evident in the case of participatory GIS projects such as GroundTruth's MapKibera, which enlisted residents of Nairobi's largest slum to map their neighbourhoods and claim public services for the first time (Berdou 2012). However, the participatory approach taken by organisations such as GroundTruth and Ushahidi (ibid.) has not so far become connected with the high-level institutional interest in data science using digital data from LMICs. Although it has been argued that participatory methods produce more accurate data on the micro-level (Chambers 1997), there has always been a deep divide between the participatory approach and the more top-down approach traditionally employed by international-level development actors, which the examples outlined in this paper will illustrate.

The perceived lack of good data is one reason why researchers and policymakers, among others, are now looking towards data produced by mobile phone and other new sources of data such as social media (Kirkpatrick 2011) and online prices (the latter in the case of MIT's 'Billion Prices Project' (Cavallo 2013), as outlined below). It is thought that these may, in some cases, be more reliable than existing statistics collected by governments. They can, in many countries, offer a population-wide perspective, are produced automatically rather than being embedded in institutional practices and biases, and are usually not subject to censorship or manipulation by intermediaries for political reasons (one of the main problems cited by Jerven [ibid.] with existing statistics).

These born-digital data also have drawbacks, however. First, access must still be negotiated or bought, which potentially means substituting negotiations with corporations for those with national statistical offices. Second, the meaning of such data is not always simple or stable, and local knowledge is needed to understand how people are using the technologies in question. Third, bias in proprietary data can be hard to understand and quantify (Gonzalez Bailon et al. 2012). Furthermore there are risks to privacy in the absence of a clear ethical framework or set of rules for handling and sharing 'born-digital' data: anonymisation techniques are unreliable (de Montjoye et al. 2013); there is less awareness in LMICs of the implications of making personal data public, and digital data protection is not yet a concern for a majority of LMIC governments (Greenleaf 2012). This paper will explore these issues in greater detail with the case studies below. Of course, some of these issues apply to high-income countries (HICs) too, so one of the aims of this paper is to highlight where there might be differences between the two sets of countries.

In the light of such privacy concerns, it should be noted that this paper specifically explores the questions surrounding data which is emitted as a byproduct of the uses of communications technologies, i.e. data which is emitted rather than consciously transmitted by the user. As such, we do not attempt to cover issues to do with Volunteered Geographic Information (VGI) or other forms of consciously volunteered data, such as crowdsourced data collected through dedicated platforms such as Ushahidi. We are interested here in data which is not submitted for a particular purpose by users who are aware that they are making the submission, but in the wealth of other data 'in the wild', which are increasingly being used to inform development policies and interventions. These data are interesting for several reasons, one being that they are particularly amenable to usage practices which aim to classify and sort populations based on needs and access to resources. These practices undoubtedly blur the boundary drawn by Lyon (2007) between 'care' and 'control'. As Lyon argues, however, the ubiquity of data production thanks to digital technologies suggests that rather than centering on Foucault's notion of the hostile Panopticon, researchers must take account of a broad range of objectives and methods involved in watching human activities, including various types of surveillance which are directed toward 'care and protection' (ibid: 67). The boundary between surveillance and development is further blurred by the problem that in most LMICs legal frameworks for data protection, and civil society activism about protecting personal data, have yet to emerge (Greenleaf 2012).

This paper focuses on the use of 'big data' deriving from digital communications technologies in development policy and planning. Development itself is a contested term which has been characterised by a huge range of understandings (e.g. Lucas 1988; Chambers 1997; Collier 2007) and of critiques of these understandings (e.g. Sen 1999; Easterly 2014; Sachs 2005). Here, we adopt the broad perspective on development characterised by the (predominantly international) institutions engaged in promoting the use of digital data for policy and planning relating to LMICs, which includes both economic development and facets of human development such as health and human rights. Further, we aim to distinguish the use of big data to inform development policy from the perspective of ICTs as a tool for development (ICT4D), and from the more recent M4D movement which focuses on mobile phones (Donner 2010).

We next examine how the potential benefits of knowing more about the subjects of development are being weighed against the potential risks of using, and trusting, this born-digital data. We will focus on three of the main uses of such data in the development field today. The first is mobile data as a predictive tool for issues such as human mobility and economic activity, through the 2012–2013 Data for Development project. The second case is the use of mobile data to inform humanitarian response to crises, through the Flowminder project. Third, we will explore the use of born-digital web data as a tool for predicting economic trends, and the implications these have for LMICs. We discuss how these cases relate to the larger use of digital data in research and policy analysis on LMIC development issues, and we consider the benefits and potential risks of using big data in this context. Our analysis is based on 26 interviews (for cited interviews, see list at end) conducted with project leaders and technicians working with big data on questions of policy and practice in LMICs, combined with analysis of the relevant literature. The paper is part of a larger project with a larger set of interviews (more than 100 completed so far) about the uses of big data in research funded by the Alfred P. Sloan Foundation at the Oxford Internet Institute.

Born-digital data as a policy resource

In the global North, discussions about mobile phone and online data have recently mainly been in the context of debates about privacy or the illicit monitoring of data by governments. The Snowden revelations have led to the increasing awareness that the data emitted in the course of the use of everyday technologies such as mobile phones and the internet can speak eloquently about users' lives, connections and activities. Where international development institutions such as the United Nations (UN) or Organisation for Economic Co-operation and Development (OECD) have become interested in using these types of data for research and policymaking, technical concerns rather than privacy risks and issues of bias have tended to dominate the discussion to date (notably at the Internet Governance Forum in Bali, in October 2013, and at the Data for Development conference within the Netmob conference, in May 2013. The latter is dealt with in greater detail below). It is possible, however, to look to the literature on technology adoption and its social consequences to find the most important topics likely to arise as this use of new data sources progresses, which can be summarised under three headings. First, understanding how the social and ethical dimensions of digital technologies impact their use and thus their utility as a source of data for policymakers, including the risk of technology becoming a factor in socioeconomic inequality (Heeks and Kenny 2002). Second, the politics which arise around data as a form of political representation, where new technologies create winners and losers in terms of political influence. Third, the question of whether the rise in importance of these new sources of data is at the expense of other types, or whether complementarity will be sought between nationally gathered survey data and born-digital datasets.

Much research has emerged on mobile phones' adoption and potential uses in LMICs, along with their social consequences, which may be used to broaden the discussion of how data deriving from them may be useful in thinking about international interventions for human or economic development. The gathering of 'emitted' big data from LMICs is a fairly new phenomenon given how recently mobile phones, in particular, have become widely adopted. The use of such data to inform interventions began with humanitarian response (perhaps most notably for the second case study below), and some widely publicised successes in the humanitarian sphere have led to discussions of big data as a resource for broader policy research. At present there are several initiatives active within large development organisations including the OECD, World Bank and UN which are seeking access to an ever-greater cross-section of big data.

The potential utility of big data as a lens on LMICs is demonstrated by the high rates of adoption, but closer study reveals that the demographics of use are complex. As Doron and Jeffrey (2013) point out in their study of India, mobile phone use is highly differentiated by gender and income level. Consider, as documented by Doron and Jeffrey, that new tensions may arise when, for example, a new member of the household with a phone may now have control over finances at the expense of another who does not. Or again, we can think of gender in the household, where, apart from power asymmetries, which include that women must often use the phone only in public, it is often the case that husbands use newer multimedia phones and pass the older phones onto their wives: it is likely that the two types of phones produce different types of data arising from how they are differently used. These and other complexities 'on the ground' may affect the way that data is being generated and may be used as measures (though they do not necessarily have an impact on the validity of analyses which use mobile big data).

Even these complexities need to be put into the larger context of one of Doron and Jeffrey's main conclusions, which is that a major impact of the mobile on Indian society is that it 'drew India's people into relations with the record-keeping capitalist state more comprehensively than any previous mechanism or technology' (2013: 224). The links to Jerven's point about the importance of being counted are clear: states are able to govern, tax and control to the extent that they can identify their citizens, and censuses, taxation, health and a myriad of other record-keeping functions are priorities not just for India, but for all functioning states. Hence it is worth highlighting that data, which may seem to be innocuous, can have major societal repercussions [as Scott (1998) has argued].

A different way to stress these complexities is by pointing to the role of data in development politics. Jerven (2013) has documented how data are politicized, and that even formally 'correct' counts may not be accepted. He offers the example of Kenya, where censuses have repeatedly been rejected by sectors of the population who feel underrepresented. Counts, he notes, must be agreed upon, disseminated and accepted as legitimate by the population at large in order to be useful to country authorities as support for policy decisions. These questions of the politics and credibility of data are not exclusive to LMICs. Numerous HICs have seen popular mistrust of the census [such as the Netherlands (Blessing 2005) and the US (New York Times 2000)], and several countries are currently considering replacing or at least supplementing the census with new types of born-digital data to save money and to attempt to remedy popular mistrust (Keeter 2012). It is important to stress that big data is currently mainly informing policy on the part of high-level institutions such as the UN and OECD which can muster the technical and analytical resources to work with very large, unstructured datasets. So far, our research has identified no channels linking the multilateral to the country level, through which the use of big data might be democratised (our first case study is evidence of this dynamic).

As Taylor (2014) argues, there are a number of problems associated with the greater visibility that big data brings in the context of LMIC societies; foremost among them that the ethical, legal and social frameworks to do with data protection and privacy that exist in HICs do not exist in LMICs. However, the numbers of adopters of mobile phones and the internet suggest that big data will become an inescapable element of evidence-gathering for policymakers: the citizens of LMICs are soon to be the largest producers of digital traces: as a group, these countries now have 89 % mobile penetration and 31 % internet usership (ITU 2013a), and are forecast to provide the majority of geolocated digital data by 2020 (Manyika et al. 2011).

Big data and digital traces (specifically those categorised as 'observed' or 'inferred' data' by Hildebrandt (2013) in contrast to volunteered data where the subject is aware of their data emission) have prompted various institutional responses in the development community. These come mainly from multinational institutions such as the United Nations, which has set up the Global Pulse initiative to organise the sharing of digital data from LMICs worldwide, and to operationalise the idea that 'shared data constitutes a public good' (Kirkpatrick 2011). For-profit and nonprofit intermediaries such as Jana (http://www.jana.com/) and Flowminder (discussed below), are also evolving to broker between corporations and data users. So far, high-profile examples of the use of new forms of digital data in development have included Twitter for epidemiology (Chunara et al. 2012), internet content mining for global pandemic disease forecasting (Wolfe et al. 2011), and online platforms for reporting election irregularities and violence via mobile phones (Ushahidi 2012).

Research using mobile traces is at the forefront of these changes in what is seen as 'development data' (though we will also discuss web-based data below). This form of data is increasingly shaping perceptions of population presence and movement in countries where real-time data is hard to acquire. For the first decade of the 2000s, the research frontier was mainly characterised by studies of human mobility (for an overview see Blumenstock 2012), evolving from industrialised-country studies which demonstrated that mobile calling data could be used to model and forecast mobility (e.g. Eagle and Pentland 2006) to studies modelling mobility in LMICs (e.g. Eagle, de Montjoye and Bettencourt 2009; Frias-Martinez et al. 2010). Mobile data on LMICs, however, has been sparse in comparison to the data available to researchers on industrialised ones, possibly because network operators have been wary of releasing data in regions with few data protection regulations.

Given the importance and accessibility of mobile phone data for informing international interventions in the development field, we next explore three case studies, two of which are based on this type of data, and one on internet content, the other growing source of digital information on LMICs.

'Development' data: Orange's D4D challenge

Against this backdrop, we can turn to our first case, the 'Data for Development' (D4D) project run by mobile operator Orange in 2012–2013. This research challenge considerably advanced the field with a controlled release of 2.5 billion call records from Orange's Côte d'Ivoire (Ivory Coast) subsidiary, with the aim of having researchers 'help address the questions regarding development in novel ways' (Orange 2012). This was the first such release to be specifically termed a 'development' project, and was consequently endorsed by the United Nations, the World Economic Forum and a host of high-profile academic institutions including MIT and Cambridge University (ibid). The dataset, comprising data from around five million users, consisted of four elements: records of 2.5 billion within-network calls and SMS exchanges over the period of a year; the spatial trajectories of 50,000 users with high resolution over a period of 2 weeks; the trajectories of 500,000 users at lower resolution over the course of the year; and communication subgraphs showing the communication networks of 5,000 users over the year. The dataset was released through a formal application process to 250 teams of researchers worldwide who signed an agreement with Orange not to distribute the data further, and who then received datasets which were anonymised to international standards by Orange's technicians.

The D4D project brought mathematicians, statistical physicists interested in complex systems, European transport policy experts, and other data scientists into a project which was nominally about guiding development policy for one of the world's lowest-income countries. Yet D4D also illustrated that the researchers with the skills to understand and analyse a dataset may not always connect with the researchers and policy experts who know the questions to ask 'on the ground' about the benefits. National authorities were not involved in conceptualising the project's development aims, and only one of the 250 research teams who received the data visited Côte d'Ivoire. One geographer was asked to comment at the presentation of the results,[2] but the main point that was made was that the project was not designed to have relevance to the particular conditions in Côte d'Ivoire: instead, it was designed to attract the best data scientists with an unusual dataset, and to motivate a new perspective on their analysis by encouraging them to relate their findings to 'development'.

This framing of the project raises the crucial problem of what 'development' means to data scientists, and how that determines what data science can achieve within the field of international development. Without research questions that are relevant to the country context, researchers are less likely to produce findings with any local impact. The project's directors acknowledge this problem: Nicolas de Cordes (de Cordes 2013) describes the challenge as having 'a fuzzy objective' in terms of its understanding of development, while Vincent Blondel (Blondel 2013) explains that the project originated in the desire to do something 'big and interesting' that went beyond the 'churn prediction' (prediction of how often mobile customers switch services) that had been the topic of past challenges.

The project resulted in 74 peer-reviewed papers on human mobility, population estimation, economic activity estimation, data mining and public health (Netmob 2013). Among the products of the work were new data mining techniques and novel ways of mapping communication networks with mobile data. However, the project was, and remains, fairly disconnected from Côte d'Ivoire and its policymakers. The plan remains to organise a 'feasibility event' to discuss the results with groups in Côte d'Ivoire, but at the time of writing this has not occurred. Part of this disconnection relates to nervousness about whether the project's findings will be politically contentious, since the data clearly identify the ethnic and spatial characteristics of communication networks and mobility patterns during a year when the country was undergoing a civil war. Blondel relates:

> These mobile phone datasets, they are real goldmines and in particular if you have the full datasets they can tell you so much it's amazing. … there are good things that can be done with and also bad things that can be done with it; this is why the datasets produced for D4D have been carefully prepared. (Vincent Blondel 2013)

Blondel's comment raises some of the central questions involved in the use of big data to inform development. First, the security of the data. Orange invited researchers to 'attack' the dataset in order to check that the anonymisation had been successful. A subsequent study (forthcoming by Sharad and Denezis) shows that anonymization can be made more effective, but point out that anonymisation can never be perfect as it depends on future rather than current challenges to the dataset, including linking and merging using datasets which have yet to be collected (ibid). The project's ethical framework was also hard to determine: D4D organisers found themselves faced with a lack of national or international regulations or ethical frameworks with regard to the privacy of data subjects, or the subsequent use or sharing of the data:

> It happened that Ivory Coast is in a region where several countries have defined a formal approach about privacy, but Ivory Coast has not signed this agreement. (Nicolas de Cordes 2013)

It was thus left up to Orange to self-regulate with regard to data protection. Although the project had to ensure that it complied with the highest standards within the European Union about privacy and Orange also checked that it would comply with all legal requirements, such ad hoc ways of dealing with data regulation (or rather, the absence of established procedures) are far less than optimal for carrying out research in this area. It can be mentioned that, apart from the fact that a civil war may add additional sensitivities to location-based data, the ability for commercial operators to relate communication networks to mobility patterns without explicit consent is not unique to research in Africa: indeed, a number of

[2] Results from the project were presented and discussed in a dedicated session at the NetMob 2013 conference, held at MIT, May 1–3, 2013.

studies have been done, for example, in France and Norway, which have involved similar types of analysis (Licoppe 2004; Boase and Ling 2013). Beyond the need for effective data protection regulation in LMICs, there is clearly a case to be made that frameworks should be developed that can inform research standards, government and corporate use of mobile data. Such frameworks would go some way to addressing the inevitable power asymmetries of data science conducted on data from LMIC subjects by scientists in remote locations, although these asymmetries extend beyond the legal aspects of data protection, and clearly merit further research.

As with any research in the field of international development, the data scientists were at risk of not fully understanding the local context. For example, the winning paper in the 'development' category of the challenge (Berlingerio et al. 2013) provided a data-analysis tool to optimise public transport efficiency in Abidjan, the capital city. However, they could only access data on the formal transport system, which covers only 10–30 % of Abidjan's city transport (Lombard 2006), with the rest handled by small-scale informal operators. Similarly, the subscriber base of Orange in Côte d'Ivoire comprises five million cusomers out of a population of 22 million, implying an unknown sampling bias. Thus the analysis in the paper applies what is probably a non-representative sample of mobile phone users to draw conclusions about what is certainly a non-representative sample of the city's transport dynamics. The paper presents an analysis which addresses the problem of an inefficient transport network, whereas the actual conditions are of a highly responsive, mixed formal-informal system where the main constraints are constituted by the city's traffic capacity (Godard 2003).

Other issues arose around the competitive advantage which might derive from access to the company's data on subscriber location and activities: de Cordes noted that it was important in preparing the dataset to avoid "giving precious information to our competitor on the market share at the antenna level" (de Cordes 2013), and checks had to be conducted that none of the research teams applying for data access were from competing network operators. Furthermore, and reminding us of Jerven's arguments, de Cordes notes that per-antenna data can provide a proxy for GDP with a high level of accuracy, but that this raises political issues on the international level that the number of calls made could be used as a proxy for GDP measures.

> One research team estimated the GDP level at antenna level, and we have 1,300 antennas in Ivory Coast, so that makes a very, very fine visibility on the development of Ivory Coast, which can then rearrange the funding of the World Bank, of the United Nations … But this is a very sensitive message, because those might not be the official statistics of Ivory Coast, which are sometimes several years out of date… (Nicolas de Cordes 2013)

This illustrates how a mobile dataset can potentially be politicised on at least three institutional levels, each with different practical implications: the national level where governments may or may not have access to data deriving from their citizens, the corporate level where companies must evaluate the political risks of analysing their data, and the level of international development organisations, whose role as data analysts positions them between these other institutions. Moreover, the D4D experience demonstrates how difficult it may be to bring data science and development together when the data relates to very low-income or fragile-state contexts: while researchers may be strongly motivated by the desire to contribute to development in some way, producing high-quality scientific conclusions does not automatically do so. In fact, the gap between academia and public policy, which is often wide, is especially so in lower-income countries where there is often no access to academic publications, even if policymakers were aware of the project and wanted to read its results. Turning research into policy influence takes effort and specialist knowledge even in HICs, and in the case of D4D, the suspension of any plan to put the findings to use in Côte d'Ivoire implies that due to potential political issues with the data, the project may have made more of a contribution to data science than to development.

Crisis data: Flowminder and data as a tool for epidemiology

The use of new digital datasets for humanitarian purposes brings up different, and possibly fewer, problems. The urgency of responding to a crisis presents a utilitarian argument for making data available as fast as possible, and for sharing it only with those involved in the response. Flowminder, a project led by Linus Bengtsson of the Karolinska Institute in Stockholm, was formed in response to the need to predict movement after the Haitian earthquake and subsequent cholera outbreak of 2010 in order to control the spread of the disease (Bengtsson et al. 2011). Bengtsson's study of the Haitian project received widespread media attention [among others, in the New York Times (2011) and the BBC (2011)]. In this case, rather than using call data as in the D4D dataset, the researchers gained access through phone companies to the location of phones as transmitted by SIM (Subscriber Identity Module) cards to the cell towers. Using 1.9 million signals from 42 days before the earthquake and 158 days after, the researchers used the number of signals to extrapolate to the number of people moving[3] out of Port-au-Prince, potentially bringing cholera with them. They then used this mobility data to identify areas outside the city at risk of cholera outbreaks as a result of the out-migration (Bengtsson et al. 2011).

The study demonstrated that 'routinely collected data on the movements of all active SIM cards in a disaster-affected nation could, with potentially high validity, be used to provide estimates of the magnitude, distribution, and trends in population displacement…[it also]…found that the method was feasible to use for close to real-time monitoring of population movements during an infectious disease outbreak'(2011: 7). The study was validated by comparing its results with those of the National Civil Protection Agency (which counted buses and ships leaving the affected area) and with a household survey of a representative sample of the population carried out by the United Nations Population Fund. The accuracy of the Bengtsson et al. results compared favourably with the estimates of the National Civil Protection Agency (NCPA) and were 'similar to' (2011: 4) the estimates to the United Nations Population Fund survey (UNFPA).

At least part of the success of the Flowminder project comes from its not-for-profit status and aims. The organisers have been approached by private-sector entities to collaborate, but have not done so because they see their access to data as predicated on remaining politically and economically neutral:

> We see it as quite dangerous for our credibility if we were to mix the two [nonprofit and commercial]. The mobile phone world is an extremely competitive arena, especially between operators, and we need to be very, very neutral, so we decided as a principle to not accept monetary support from operators. (Linus Bengtsson 2013).

If data science is to be brought together successfully with humanitarian response, timely data access is critical. Bengtsson et al.'s Haitian study is superior in two respects: one is that the data could, in principle, be obtained for continuous and extended periods and in near real time (or close to when the data were obtained from the mobile operator), which was 12 h later for the cholera outbreak (in fact, data were obtained for a period adding up to 10 weeks related to the earthquake and 8 days for the cholera outbreak). Second, the data were readily available, whereas the collection of such data by other means, including by the NCPA and UNFPA whose data were used for comparison in this case, is a very labour-intensive and costly undertaking.

One drawback of the use of mobile data for epidemiological purposes, on the other hand, is the representativeness of the dataset: if, for example, the mobile operator who has donated the data is favoured by richer mobile phone users, this represents a systematic bias which affects the data's ability to predict population movements. The researchers in this case were lucky that Haiti has two mobile operators and they obtained data from the largest one, Digicel, which covers 90 per cent of the population, noting that 'we have not found evidence of differences in the companies' subscriber populations' (2011: 2). These circumstances are quite unusual since in many countries, there are more than two operators, and the subscriber populations are often quite different because the subscription packages are also different for various socio-economic groups who can afford to pay different amounts for mobile uses (or who use cheaper texts rather than voice, or prepaid plans, see Ling and Donner 2009: esp. 49–60). Other potential biases can arise from the distribution of usership: not everyone uses mobile phones, with usership particularly low amongst vulnerable and 'hidden' populations

[3] Unusually for this type of mobile phone data study, the researchers took account of the fact that in low-income countries mobile phones are often shared, and calculated the number of people per SIM before extrapolating mobility statistics.

such as children, the elderly, the poorest and women. Bengtsson shows by comparison with UN mobility statistics that for the Haitian population in question, 'users and nonusers of mobile phones had similar movement patterns' (2011: 5–6), though the bias nevertheless requires checking. And again, these conditions in Haiti are unusual; most other places are bound to have more difficult biases to assess, as demonstrated by the D4D case above. The use of mobiles for tracking population movement in disaster relief or in other settings may become more commonplace despite the many associated limitations because the data are readily available and the benefits arguably outweigh the costs. In any event, having demonstrated the feasibility of this method, it is likely to be extended to other situations. For disasters and relief operations, it will no doubt be argued that utilitarian considerations override other considerations, such as potential privacy risks and the problematic nature of the data. Klein (2008) has shown how crisis can provide a context where policy change may go unchallenged due to weakened civil society and political institutions. Her research underlines the importance of applying and enforcing an ethical baseline for data sharing and use, and for mechanisms to regulate cases where the use of personal data may need to go beyond established boundaries.

So far, no inter-regional or global frameworks for data protection or privacy have been developed, with regulation remaining either national or confined to regional blocs such as the EU. The use of extensive datasets produced in LMICs, though with analysis conducted in HICs, has as yet not been addressed by regulators and it is uncertain where responsibility for such regulation should fall. This is a particularly thorny issue given that multilateral institutions such as the UN and OECD are now stakeholders rather than neutral parties, due to their advocacy for the use of big data to observe conditions in LMICs. Although there is an indication that organisations such as the World Economic Forum and Privacy International are attempting to tackle this question, the only solution has been, as shown by the Orange case study above, that companies involved in such data transfer and analysis should self-regulate based on regulation in their home region. If more wide-ranging regulation can be developed, however, ethical and normative frameworks for digital data sharing in cases of crisis could temper a utilitarian approach with a more consequentialist perspective, since once data is shared, it becomes hard to regulate who may acquire and use it, and for what purposes.

Macro-economic data: tracking prices online

Our final example of big data in development comes from the Billion Prices Project (BPP) at MIT. This project was initiated by Alberto Cavallo at MIT's Sloan School of Management, (and now co-managed by his colleague Roberto Rigobon) to challenge what he felt to be a misleading inflation index managed by the Argentine government. The story of the BPP reflects some of the issues brought up by Jerven in his critique of the way country statistical offices are affected by political conditions and demands.

> Back in 2007… the government of Argentina, where I am from, started having problems with inflation rates, so they decided after the controls and other techniques that didn't work, they decided they were going to intervene at the statistical institute and they essentially fired everyone that was responsible for building the price index. And weird things started happening to the data, the inflation rate stabilised, people started saying that the data was not reflecting the actual inflation rate in the country. (Alberto Cavallo 2012)

Cavallo decided to create his own inflation index as a comparator for the official one. He programmed a web scraper to find prices for everyday goods posted on the web by the country's supermarkets, and compiled an index from the trends in those prices. The data is entirely public, and the costs involved in the BPP come mainly from customising and monitoring the software created by Cavallo, so that it can scrape data accurately despite changes in the websites' structure. This big data approach is cheaper financially and in terms of labour than compiling the official index: the coding for the original index was done by Cavallo alone, and although the project's success has led to it adding staff in order to analyse more countries, in theory such an index could be compiled by a single researcher from his or her desk.

The BPP has been influential, first because it produced what seemed to be an inflation index that was more intuitively reflective of perceptions in

Argentina than the government's, thus adding weight to the suspicion that the statistical agency was under political pressure (Financial Times 2013), but also because it demonstrated that big data could provide a viable alternative perspective on economic trends under the kinds of fiscal instability that challenges the collection of accurate statistics. It is now available for multiple countries, is updated daily, and can focus on countries where official statistics may be problematic. These aspects make the BPP an example of the way that big data could lead to a rethinking of development statistics for other countries. However, location may play a role in determining whether this type of data can gain public trust and political traction. Argentine citizens have comparatively high internet usership among Latin American countries (9.6 % overall, but 46 % in the capital region (ITU 2013b)), and a relatively free press, ranking fourth in Latin America for press freedom (Reporters Without Borders 2013), so that citizens are likely to be able to compare official statistics with the BPP's index. Moreover, unlike population or migration numbers, inflation has an intuitive aspect where consumers can tell if they are spending more on necessities than previously, and whether the government's estimate of how much agrees with their own. This index therefore has political traction which may be harder to achieve in lower-income, less technology-literate countries—suggesting that place plays an important role in the credibility of new types of data.

Cavallo subsequently further developed the index to be used as a food security indicator applicable across a wide range of countries. The most recent application is a comparison of the post-earthquake behaviour of food prices in Chile and Japan (Cavallo et al. 2013), which shows that the countries addressed food security in very different ways. The study of these countries' different price trends demonstrates the complex dynamics of supplier-retailer-customer relationships during economic shocks, offers clues as to how retailers make decisions to raise prices or not under those conditions, and makes it possible for policymakers and humanitarian organisations to gauge more accurately how to respond to food insecurity in crises. As online food prices become more available for LMICs, indices such as Cavallo's may gain in importance as a way of triggering international response to food security problems. With this research, we can also see how Cavallo's use of big data relates not only to the economics of prices, but also to the development policy and disaster and disease response of the other two cases discussed.

Scraping the web also has wider uses in development economics, and beyond. With the development of aggregators such as Flowminder and Jana, and data science initiatives such as Global Pulse, the lines between different data sources may increasingly become blurred, as Cavallo seems to suggest. Mobile data and social media data are natural partners, Bengtsson notes (interview, 2013): users may be tracked via their mobile, but may also be using it to access the internet and post on social media, producing multiple types of trackable data emission which provide different facets of phenomena such as financial transactions, mobility and communication networks. The analytical power of these datasets will increase where they can be brought together to inform a particular question. Robert Kirkpatrick, director of Global Pulse, notes that this variety of data sources is central to the effectiveness of big data as a development tool:

> This has applications in human rights, it has applications in disaster response, in disaster resilience and disaster recovery. And we have moved, we have shifted away from kind of looking at what we might call traditionally strict development issues, to a framework that is much more wellbeing centred. I mean, we are basically saying how do we monitor human wellbeing in real time? (Robert Kirkpatrick 2013)

One issue that is bound to become pressing in the effort to monitor issues as complex as inflation or wellbeing is that big data are often not just relatively cheap but, in relation to problems which conventionally require massive surveying capacity to resolve, also relatively easy to analyse if the right technical skills are available. As Robert Kirkpatrick puts it, one can do 'passive monitoring at a very low cost'. This passive monitoring can lead to new ways of looking at concepts within the development field such as resilience: Kirkpatrick asks, 'can we understand economic vulnerability, by studying changes in mobile phone usage and air time purchases, in response to different contextual changes?' (Kirkpatrick 2013). He also describes a Global Pulse project where air time purchases are used in combination with other continuously updating indicators such as 'social network dynamics and movement patterns' (Kirkpatrick 2013) to research the way in which individual communities return to normal after a shock such as an earthquake. This suggests the objective is to create a large-scale, longitudinal database which can be continuously mined for information on aspects of human and economic development—an observatory of poverty, resilience and growth.

Such an observatory has much to offer LMICs where vital data are scarce. It may even create a temptation to use big data in the place of data that are more difficult to gather, as for Cavallo's Billion Prices Project. Importantly, this view of development also extends beyond LMICs. While the idea of using web-scraped price data from supermarkets emerged because Cavallo could not find the necessary data for his countries of interest, it applies equally to the US where high-quality data have been gathered by hand—researchers going into supermarkets with clipboards—for a long time. In this case too, arguments are emerging about the costs and benefits of checking prices manually, versus doing this by scraping the web—which may introduce inaccuracies—but these may be overridden by the sheer volume of data that can be analysed.

Research is needed into the nature of bias in scraped data, particularly in order to understand whether errors are randomly distributed or systematic. At the moment the answer to this question is unknown, largely because computer science has not focused significantly on issues of bias in datasets. This has so far been the territory of social science (e.g. Gonzalez Bailon et al. 2012), and even there is at an early stage. Resolving these issues may require greater collaboration between social scientists and data scientists, which is largely an issue of institutional support. (It can be mentioned that Cavallo, too, like Bengtsson, compared his scraped prices with manually collected data, and found 'a remarkable similarity…in the timing and size of price changes between online and offline samples'(2013: 3). Moreover, scraped supermarket prices in Latin America are 'not a major issue' according to Cavallo because they represent a large proportion (over 40 %) of what goes into the consumer price index (2013: 8)). However, in the development context, where there may be greater resource constraints, these cost/benefit arguments might play out differently. At the same, such potential broader and more widespread uses of big data also compound and complicate the regulatory issues, the issues of access to commercial sources of data and details of their origins, and others that we have discussed.

Discussion

The data involved in the case studies above can be distinguished from volunteered sources of data about LMICs (Volunteered Geographic Information and other crowdsourced data), and have, in comparison to these, received significantly less attention in both scholarly research and in the press. The case studies offer some clues as to why this may be so: first, they offer examples of data emitted entirely under corporate auspices rather than through nonprofits such as community mapping projects or authorities involved in crisis response. Second, the emission and use of these data generally occur without the awareness of data subjects, making them again different from crowdsourced datasets. Third and finally, these data offer an unprecedented level of detail—albeit less clear and structured than volunteered data—on micro-level activities and transactions.

These characteristics give rise to different epistemological concerns from those raised by volunteered data. In the case of volunteered data verification tools are being developed to answer a clear need for greater certainty about the origin and veracity of social media in particular (Schifferes et al. 2013). In contrast, the big data in two of our three case studies comes from a single corporate source and is released under restricted conditions, making verification through replication difficult. (In the case of the Billion Prices project, multiple corporate sources at least make comparison possible.) This characteristic of limited replicability and thus verifiability is important when considering the politics of data as representation, as discussed above.

The second major issue which arises with this type of corporate data is that of unknown bias. This problem is not particular to mobile or price data, but also spans social media such as Twitter where users are not representative of the population at large. It also relates to web-content datasets that are not commercial and fully available to researchers: for example, Wikipedia editors are also not representative of the population at large (Crampton et al. 2013). This is a problem which may be viewed through the lens of technology adoption, since different groups will adopt new means of communication and information-seeking at different rates and in different ways (Musa et al. 2005). Understanding the local conditions for adoption and use is therefore important in order to understand what a particular dataset can tell the researcher—but this kind of qualitative inquiry is not usual amongst the disciplines from which data scientists tend to be drawn such as computer science, informatics, or even theoretical physics [as demonstrated by the participants in the 2013 Netmob subconference on Data for Development (Netmob 2013)]. This issue of bias adds to the potential political implications of using big data to profile and intervene in LMIC populations—however, in the case of big data these implications may be invisible to data subjects. This explicitly contrasts with volunteered information: with the big data described here, data subjects are unlikely to be aware of the connection between their data and any resulting policy, whereas the data subject who volunteers data does so with a specific goal in mind which is, at least in most situations, verifiable.

The issues of privacy and data protection have both been highly publicised in this context, due to the lack of effective regulation of the data collection and analysis discussed here. In contrast to volunteered information, big data deriving from sources such as mobile communications and internet use are extremely sensitive with regard to the protection of data subjects' information, particularly when those subjects live in places of limited statehood where enforceable regulation is often absent (Greenleaf 2012). The stated aim of the United Nations Global Pulse to build an 'observatory on poverty'[4] may be in conflict with the right to privacy of LMIC citizens, which remains as yet unarticulated in legislation and is thus unenforceable. What constitutes the ethical use of data emitted by LMIC citizens will be partly determined by the development of concepts of privacy and rights over one's own data, which have not so far been encoded in law in most LMICs. They will also, however, be determined by the precedents currently being set in terms of access to LMIC data under the rubric of 'development', given that collaborations, data access agreements and analytical practices are now being established and solidified without consultation with LMIC governments or data subjects.

Conclusion

We began by outlining why the use of big data in the development field is set to increase exponentially as new data sources become accessible. We showed at the hand of three cases that where it is used to inform development policy and humanitarian response, big data raises some particular questions. First, can it remedy the relative lack of information available to policymakers regarding LMIC populations and trends? Second, can the current problems of privacy, proprietary access, uncertain bias and a lack of accompanying qualitative information be resolved adequately, so that big data tells a comprehensible and verifiable story? And last, how should policymakers weigh the risks and benefits of big data as a policy tool? These issues seem to lead back to the question we posed throughout this paper: what does it mean for data scientists to be involved in development research? Data science conducted with the aim of informing development policy must, by definition, involve an understanding of the policy area in question, and importantly the analysis must be combined with understanding of the local context. Without these characteristics, research only informs the field of data science rather than development policy.

Regarding the first question posed above, we have shown here that big data do not *per se* represent a solution to the lack of population-level information on LMICs as idenfitied by Jerven (2013). This is because the real challenge is not simply gaining more data, but gaining data which is relevant to country priorities and can gain traction in terms of supporting policy and lead to more informedness and greater transparency. This last, as Jerven has pointed out, is a greater challenge than even creating 'good' statistics. Useful data is the data that is useful not only to international experts but also to country policymakers, and can thus become a tool for consensus-based country-led development. There are two sides to this challenge, both of which reflect larger structural problems in the development field. First, efforts must be coherent and coordinated: various international agencies such as the UN and G20 think of big data as global 'early warning systems', or as ways to address global crisis issues, yet there are

[4] Robert Kirkpatrick, director, UN Global Pulse, in speech to the Internet Governance Forum. Bali, October 23, 2013.

already multiple overlapping and competing projects and little coordination between them. Also, LMIC policymakers must be involved in the identification of ‘development problems’, and must have access to the data once it becomes available in order for such data to have political acceptability and buy-in. For instance, will a government want an early warning system which determines aid flows, media attention and international action—but without control over the data? Any such effort, if it arouses this sensitivity, must ensure that it has the relevant consent from countries if they are to be ‘on board’.

Regarding the second question posed above, big data’s privacy risks and questions about its reliability may perhaps be seen as two sides of the same coin. According to our interviewees, detail, richness and multidimensionality are characteristics of big data which can go some way towards resolving its reliability concerns. If researchers have unrestricted access to the dataset for resolving replicability questions and related qualitative information to both understand issues of bias and to increase the data’s relevance as a policy tool, big data may be an unprecedented resource for informing policy interventions. However, this is a big ‘if’ because these conditions have not been achieved in the case of traditional datasets, let alone in the data-science-dominated process of big data analysis. Although corporate ownership of the data seems a challenge in terms of unrestricted access, gaining the relevant qualitative information to inform the research is an even greater one. The convenience of conducting data analysis on a huge scale in the comfort of the offices of international institutions may be outweighed by the damage which can be done by underinformed quantitative research and resulting policy interventions. The human toll of policies such as structural adjustment or forced rural displacement in Africa (Scott 1998) illustrates that this problem is neither new nor confined to big data.

However, the ideal balance of big data and qualitative information seems to raise serious questions about privacy. Because the big data discussed here is often a byproduct of individual activities, communications and transactions, this suggests that more detailed data, with wider access provisions, combined with more on-the-ground information about the populations in question may create a situation where LMIC citizens are unwittingly placed in a panopticon staffed by international researchers, with no way out and no legal recourse. As the right to privacy may not be formally framed through legislation in LMICs, data access, volume and verifiability is likely to win out over the concept of privacy, which remains a hazily defined and negative right even in jurisdictions with enforceable regulation such as the EU and USA. If the conflict between big data visibility and the right to privacy cannot be resolved in these regions, it seems unlikely that LMICs will lead the way.

We have argued here that big data provides a new type of visibility, both technically and ethically complex, to the subjects of development policy. The emergence of big data has not only made individuals more visible to policymakers, but to also to corporations. This duality is not confined to LMICs, but is arguably made more intense by the higher stakes of becoming newly visible in areas of limited statehood and potentially untrustworthy governments. Conversely, of course, there may be just as many risks involved in becoming visible to international corporations without the knowledge or protection of national authorities. Greater visibility of populations can have positive effects, as when GDP is measured more accurately, or the spread of disease tracked faster, or relief brought to disaster areas where needed. On the other hand, visibility can lead to marketing that pesters users, fleeing populations can be targeted by militaries that pursue them, or visible populations may be favoured at the expense of less visible ones.

In short, visibility has costs as well as benefits. As argued by Dandeker (1990), with the development of the modern state and the increasing provision of health, education and welfare for citizens, the state also has a need to capture more and more information about citizens in order to provide these services or benefits. Citizen rights thus move in lockstep with citizens’ obligations to divulge information about themselves. With the growing uses of big data, however, this obligation is often being translated to the private sector, with users expected to consent to broad and sweeping agreements for the collection and processing of their data in addition to paying for services. We have explored some of the (re)uses of big data which epitomise this revised social contract, and pointed out some ways in which it may cut deeper in LMICs than HICs, given that the stakes may be higher there. When agreeing to share your data may make you visible to rescuers in the case of an emergency, who would refuse? It can be argued, however, that if this new visibility is to include policymakers, corporate interests and others, LMIC citizens should be made fully aware of what they are signing up for when they become adopters of new technologies. This does not constitute an argument that today's citizens are living inside a panopticon, rather we would argue (as suggested above, in line with the argument of Lyon 2007) for the idea of a panopticon incorporating aims of care rather than merely discipline.

A first step toward more responsible and responsive data use in an LMIC-development context would be stronger connections between data scientists and those with local knowledge, combined with institutional support for those connections. As soon as a data source gains value by being able to influence the distribution of resources, it attracts politics and potentially discord. It is likely that dissent over big data generated in LMICs will increase, given that the awareness of its use and reuse is only starting to spread in these regions. If the power and knowledge asymmetries we have outlined here can be resolved or at least mitigated, this data has immense potential to improve the statistics available on the characteristics and needs of LMICs, and to contribute both to improved crisis response and more responsive and tailored development policy. It remains to be determined whether the use of big data can be cooperative, collaborative and responsive to local concerns, or whether it will follow the path of many technocratic innovations in the field of development by remaining remote and detached from country-level reality. Given the potential range and power of these new types of data, it is clear that data science can be made increasingly relevant to the concerns of LMICs, and that these potential asymmetries are well worth negotiating.

Acknowledgments We gratefully acknowledge funding from the Alfred P. Sloan Foundation for support for the project 'Accessing and Using Big Data to Advance Social Science Knowledge' at the Oxford Internet Institute. We are grateful to Ryan Burns and Jim Thatcher for helpful comments on an earlier version, which clarified the paper and improved several points, and to two anonymous reviewers for their comments and ideas.

References

BBC News. (2011). *Mobile phones help to target disaster aid, says study*. http://www.bbc.co.uk/news/technology-14761144.

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 Medicine, 8*(8), e1001083. doi:10.1371/journal.pmed.1001083.

Berdou, E. (2012). *Participatory technologies and participatory methodologies: Ways forward for innovative thinking and practice*. IKM Working Paper No. 17.

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., & Sbodio, M. L. (2013). AllAboard: A system for exploring urban mobility and optimizing public transport using cellphone data. In *Machine learning and knowledge discovery in databases* (pp. 663–666). Berlin: Springer.

Blessing, M. (2005). *Het verzet tegen de Volkstelling van 1971*. Historische Nieuwsblad nr. 8/2005. http://www.historischnieuwsblad.nl/nl/artikel/6697/het-verzet-tegen-de-volkstelling-van-1971.html.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology for Development, 18*(2), 107–125.

Boase, J., & Ling, R. (2013). Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication, 18*(4), 508–519. doi:10.1111/jcc4.12021.

Borgman, C. (2014). *Big data, little data and beyond*. Cambridge: MIT Press.

Cavallo, A. (2013). *Scraped data and sticky prices*. MIT Sloan Working Paper, http://www.mit.edu/~afc/. Accessed 4.10.2013.

Cavallo, A., Cavallo, E., & Rigobon, R. (2013). *Prices and supply disruptions during natural disasters*. Working Paper 19474, NBER.

Chambers, R. (1997). *Whose reality counts? Putting the first last*. Intermediate Technology Publications Ltd (ITP).

Collier, P. (2007). *The Bottom Billion: Why the Poorest Countries are Failing and What Can Be Done About It*. Oxford: Oxford University Press.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., et al. (2013). Beyond the geotag: Situating 'big data'and leveraging the potential of the geoweb. *Cartography and Geographic Information Science, 40*(2), 130–139.

Dandeker, Christopher. (1990). *Surveillance, power and modernity*. Cambridge: Polity Press.

de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports, 3*.

Donner, Jonathan. (2010). Framing M4D: The utility of continuity and the dual heritage of "mobiles and development". *Electronic Journal on Information Systems in Developing Countries, 44*(3), 1–16.

Doron, A., & Jeffrey, R. (2013). *The great Indian phone book: How the cheap cell phone changes business, politics, and daily life*. London: Hurst&Co.

Eagle, N., & Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing, 10*(4), 255–268.

Eagle, N., de Montjoye, Y. A., & Bettencourt, L. M. (2009). Community computing: Comparisons between rural and urban societies using mobile phone data. In *International* *conference on computational science and engineering, 2009 (CSE'09)* (vol. 4, pp. 144–150), IEEE.

Easterly, W. (2014). *The Tyranny of Experts: Economists, Dictators, and the Forgotten Rights of the Poor*. New York: Basic Books.

Financial Times. (2013). *Argentina: Questioning official inflation can land you in jail*. Accessed September 13 http://blogs.ft.com/beyond-brics/2013/09/13/argentina-inflation-diverging-from-official-numbers-can-land-you-in-jail/#axzz2gYghm6jJ.

Frias-Martinez, V., Virseda, J., Rubio, A., & Frias-Martinez, E. (2010). Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. In *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development* (p. 11), ACM.

Godard, X. (2003). *Urban transport and mobility in African cities. Crisis and inventive disorder*. Paper prepared for TRB annual meeting January 2003. http://onlinepubs.trb.org/onlinepubs/archive/am/03-2786.pdf.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2012). *Assessing the bias in communication networks sampled from twitter*. Available at SSRN 2185134.

Greenleaf, G. (2012). *Global data privacy laws: 89 countries, and accelerating*. Queen Mary University of London, School of Law Legal Studies Research Paper No. 98/2012.

GSMA. (2013). *Sub-Saharan Africa Mobile Economy 2013*. http://gsma.com/newsroom/wp-content/uploads/2013/12/GSMA_ME_Sub_Saharan_Africa_ExecSummary_2013.pdf.

Heeks, R., & Kenny, C. (2002). *The economics of ICTs and global inequality: Convergence or divergence for developing countries? Development informatics*. Working Paper No. 10a, Institute for Development Policy and Management, University of Manchester.

Hildebrandt, M. (2013) Slaves to big data. Or are we? *Keynote, 25th June 2013 9th annual conference on internet, Law & Politics* (IDP 2013, Barcelona).

ITU. (2013a). *The world in 2013. International telecommunications union*. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf.

ITU. (2013b). *International Internet connectivity in Latin America and the Caribbean*. Geneva: International Telecommunications Union.

Jerven, M. (2013). *Poor numbers: How we are misled by African development statistics and what to do about it*. Ithaca: Cornell University Press.

Keeter, S. (2012). Survey research, its new frontiers, and democracy. *Public Opinion Quarterly, 76*(3), 600–608.

Kirkpatrick, R. (2011). *Data philanthropy: Public and private sector data sharing for global resilience*. http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience.

Klein, N. (2008). *The shock doctrine: The rise of disaster capitalism*. New York: Metropolitan.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine, 48*(9), 140–150.

Licoppe, C. (2004). 'Connected' presence: The emergence of a new repertoire for managing social relationships in a changing communication technoscape. *Environment and Planning D: Society and Space, 22*(1), 135–156.

Ling, R., & Donner, J. (2009). *Mobile communication*. Cambridge: Polity Press.

Lombard, J. (2006). Enjeux privés dans le transport public d'Abidjan et de Dakar. *Géocarrefour, 81*(2), 167–174.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics*. *22*, 3–42.

Lyon, D. (2007). *Surveillance studies: An overview*. Cambridge: Polity Press.

Mann, L. (2013). *Blogpost on OII's Policy and Internet Blog: Big Data and Informal Economies in Africa*. Accessed 2.10.2013. http://blogs.oii.ox.ac.uk/policy/seeing-like-a-machine-big-data-and-the-challenges-of-measuring-africas-informal-economies/.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: The next frontier for innovation, competition and productivity*. Washington, DC: McKinsey Global Institute.

Musa, P. F., Meso, P., & Mbarika, V. W. (2005). Toward sustainable adoption of technologies for human development in Sub-Saharan Africa: precursors, diagnostics, and prescriptions. *Communications of the Association for Information Systems*, *15*.

NetMob. (2013). *Mobile phone data for development: Analysis of mobile phone datasets for the development of Ivory Coast*. NetMob conference, May 1–3 2013, MIT, Cambridge, USA.

New York Times. (2000). *Who lives here? Who's asking? In a Black Community, Official Mistrust Hinders Census*. Mary 16, 2000. http://www.nytimes.com/2000/05/16/nyregion/who-lives-here-who-s-asking-black-community-official-mistrust-hinders-census.html?pagewanted=all&src=pm.

New York Times. (2011). *Haiti: Cellphone tracking helps groups set up more effective aid distribution, study says*. http://www.nytimes.com/2011/09/06/health/06global.html?_r=2&scp=1&sq=haiti%20bengtsson&st=cse&pagewanted=all.

Orange. (2012). *D4D project*. http://www.d4d.orange.com/learn-more.

Reporters Without Borders. (2013). *2013 World Press Freedom Index*. http://fr.rsf.org/IMG/pdf/classement_2013_gb-bd.pdf.

Schroeder, R. (2014). Big Data: Towards a more Scientific Social Science and Humanities? In M. Graham, & W. H. Dutton (Eds.), *Society and the Internet: How networks of information are changing our lives* (pp. 164–176). Oxford: OUP.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.

Schifferes, S., Newman, N., Thurman, N., Corney, D. P. A., Goker, A., & Martin C. (2013) Identifying and verifying news through social media: Developing a user-centred tool for professional journalists. *Paper presented at The Future of Journalism Conference*, 12–13 September 2013, Cardiff, UK.

Sen, A. (1999). *Development as Freedom*. Oxford: Oxford University Press.

Taylor, L. (2014). *No place to hide? The ethics and analytics of tracking mobility using African mobile phone data*. Unpublished paper, University of Amsterdam. http://www.

academia.edu/7502204/No_place_to_hide_The_ethics_and_analytics_of_tracking_mobility_using_mobile_phone_data.

Interviews

Bengtsson, Linus. Director, Flowminder. Interviewed 16.5.2013

Blondel, Vincent. Professor of applied mathematics at the Université Catholique de Louvain and organiser of the D4D challenge. Interviewed 29.3.13

Cavallo, Alberto. Cecil and Ida Green Career Development Assistant Professor of Applied Economics, MIT. Interviewed 15.11.2012

de Cordes, Nicolas. Vice President of Marketing Vision, Orange-France Telecom Group. Interviewed 16.4.2013

Kirkpatrick, Robert. Director, UN Global Pulse. Interviewed 14.5.2013