Big Data for Positive Social Change: An LMIC Perspective

Josh Cowls
Oxford Internet Institute
University of Oxford
Oxford OX1 3JS
+44-1865-287224
josh.cowls@oii.ox.ac.uk

Linnet Taylor
Institute for Social Science Research
University of Amsterdam
1012 CX Amsterdam
+31-61-6626953
l.e.m.taylor@uva.nl

Ralph Schroeder
Oxford Internet Institute
University of Oxford
Oxford OX1 3JS
+44-1865-287224
ralph.schroeder@oii.ox.ac.uk

ABSTRACT

This paper reports on a conference that was held in order to address the challenges and opportunities of using big data in developing countries. The conference included a wide range of practitioners, policymakers and researchers. It draws together various lessons learned, including brief illustrative descriptions of six of the cases represented at the workshop, under three headings of big data uses: uses for description and prescription, uses for facilitating information exchange, and uses for accountability and advocacy. Under each heading, the paper points to benefits but also to potential constraints on using big data. A final section lays out the major obstacles ahead, and makes some suggestions for overcoming these.

Categories and Subject Descriptors

K.4.0 [**Computers and Society**]: Public Policy Issues – *ethics, privacy, regulation, transborder data flow, use/abuse of power.*

General Terms

Human Factors, Legal Aspects.

Keywords

Big data, LIMCs, social change, development, ICT4D

1. INTRODUCTION

This position paper was produced by a group of activists, researchers and data experts who met at the Rockefeller Foundation's Bellagio Centre to discuss the question of whether, and how, big data is becoming a resource for positive social change in low- and middle-income countries (LMICs). This paper draws together the contributions of participants, who are listed in the appendix.

In this paper we consider the usefulness of big data to a range of communities engaged in pushing for positive social change. These include activists on digital issues such as personal data protection and privacy; organizations working on accountability and transparency; funders aiming to promote social change and interested in using new tools and resources; researchers and policymakers working on economic or human development for whom digital data is a central resource; and researchers working with big data to inform development or humanitarian action. All these groups were represented at the Bellagio conference.

We draw lessons from real-life examples of data science which may be useful to organizations and authorities regarding how data should flow and to whom, how to protect data subjects, and how to encourage new uses of large-scale digital data in civil society. Some of the central questions that arose in our discussions were also broader and relate to way data helps us to understand and portray social issues such as inequality and representation. The inevitable categorization that accompanies the analysis of digital data tends to 'flatten out' major differences in understanding contextual factors. This applies to applies to big data uses: the way a database is formed will inevitably emphasize certain characteristics at the expense others. Thus, the power of the data scientist – the person tasked with capturing and analyzing data – is significant, as they have the responsibility to consider whether something should be measured just because measurement is possible. Moreover, information expressed as 'Data' can make social categories more salient, such that the data scientist may cement differences which were previously more fluid.

In what follows, we start by offering a definition of big data informed by the context of LMICs. Next, we draw together the myriad positive uses of big data in LMICs under four more general headings, which we supplement with concrete case studies for each. We then turn to outline some of the opportunities and challenges which are introduced by the use of big data in this context, and offer ways in which the latter might be mitigated. Finally, we conclude with an overview of possible future directions.

2. WHAT CONSTITUTES BIG DATA IN THE CONTEXT OF ACTIVISM IN LOW- AND MIDDLE-INCOME COUNTRIES?

While the corporate world's definition of big data remains fairly stable around the idea of the 'three V's' (Volume, Velocity and Variety), for the purposes of our discussion we take a broader perspective which incorporates problems of interpretation (a 'fourth V' – Veracity). Our working definition of big data refers to digital datasets of an unprecedented scale and scope in relation to a particular question or phenomenon [1]. Practically, we suggest that it may be more relevant to define big data as involving particular analytical techniques and approaches using digital data; here, in relation to societal issues. Thus big data could be seen more as a process than an object.

The data discussed in this paper includes, but is not limited to, sources such as social media, mobile phone use, digitally mediated transactions, the online news media, and administrative records. The data we are interested in here can be classified into three main classes of origin: data that is *volunteered explicitly* (e.g. social media postings, digital survey responses or volunteered geographical information for open mapping); data that is *observed* (such as transactions taking place online or mobile phone call and location records); and data that is *inferred and derived by algorithms* (this includes people's social networks, trends relating to online behavior or transactions, and economic data such as inflation trends).

The real-time nature of data is also important to the definition of 'big' because it involves a change in the characteristics of information available to those interested in the dynamics of social change. The use of new digital technologies such as mobile phones and internet-based search, communications and transactions mean that in the fields of economic policy and development, we are seeing a shift from census-based knowledge where information is collected every ten years to constantly updating data which is seldom labelled as 'development'-related, and which may be difficult to access as it flows through corporate, rather than public-sector, circuits of information.

Addressing big data as a process means that we are also interested in the activities related to cleaning, processing and storing it, and how these relate to an overall objective of promoting positive change and – as importantly – behaving ethically as researchers and activists with regard to data subjects. As one participant observed: 'we first have to tackle Tedious Data before we can use Big Data', if we want to avoid problems of reliability and replicability of analysis.

3. HOW IS BIG DATA BEING USED TOWARDS POSITIVE SOCIAL CHANGE?

In this section we group together the many uses of big data for engendering positive social change in LIMCs under four headings. Even within the parameters of the definition provided above, the uses of big data in this area would be too numerous and diverse to fit into a much longer report, let alone a short position paper. However, in our attempt to group approaches under the particular rubrics which follow, we hope to provide some sense of the potential of big data in this area, and offer a framework within which different sorts of approaches can be engaged with, evaluated, and critiqued.

3.1 Description and Prediction

As suggested in our definition, big data can offer a new depth of detail on particular issues. In particular, observed data such as call records from mobile phone network operators are playing an increasingly important role in academic, development and humanitarian research due to its ability to provide high-resolution, dynamic (in terms of time, space and coverage) data sources and methods of analysis. Researchers can combine data which have not previously been combined into higher-resolution datasets from which they can detect new correlations and surface new questions. Mobile phone calling records in particular can provide deeper predictive capacity for real-time events such as conflicts, crises and elections, as shown by the work of Flowminder (Case Study 1; case studies are referenced in Appendix A).

Of course, it is not only at times and in places of crisis that 'observed' big data can be a useful tool for the description and prediction of social phenomena. For example, the stratospheric uptake of various social media platforms offers a new source of latent big data, from which the attitudes and opinions of people around the world can be inferred, such as the work being undertaken by the UN's Global Pulse initiative (Case Study 2).

Case Study 1: Flowminder

After the Haitian earthquake in 2010, many people moved away from the capital city, Port Au Prince. Researchers asked Digicel, Haiti's biggest mobile phone operator, to share de-identified information about the cell towers that subscribers were using when making calls and distributed reports throughout 2010 to relief agencies to provide information about the distribution of displaced populations around the country. The data included the position of 1.9 million subscriber identity modules (SIMs) in Haiti from 42 days before the earthquake to 158 days afterwards, allowing researchers to compare people's movement in the days preceding and following the earthquake. The estimates of geographical distribution of people across Haiti were matched by estimates from a retrospective UNFPA study/household survey. Flowminder's approach works for crises in which there are rapid population movements. There is an estimated displacement of over 100,000 people every 2 weeks (1 million displaced on average every second month) related to natural disasters such as floods, earthquakes, hurricanes and monsoons

Case Study 2: UN Global Pulse

Aiming to take the Twitter "pulse" on global development issues, social media data analysis of millions of public tweets identifies relevant keywords, and maps them against Post-2015 priority topics to show which are being talked about the most. The time frame covers 2013 to date and updates every month. By searching approximately 500 million new posts on Twitter every day for 25,000 keywords (http://datasift.com/essence/kavcgn) relevant to 16 global development topics, the project/dashboard shows which different countries talk the most about a given topic. Filtering in English, French, Spanish and Portuguese yields around 10 million new tweets each month.

However, both these examples also serve to illustrate some of the challenges big data researchers must resolve when working in the humanitarian sphere. Firstly there are issues over obtaining access to the data. In the case of mobile records, researchers must first determine the legal process involved in accessing mobile data, which varies from country to country: there is some nervousness in the communications industry about how to respond if asked for this data, so researchers may run into significant problems where access is not mandated by law. In the case of social media data, many companies such as Twitter put severe limits on the scope of access researchers enjoy. Large non-governmental projects such as those run by UN Global Pulse can make special pro bono agreements with such companies, but replicating this type of research would take similar arrangements or the ability to purchase data from a commercial broker. The alternative is the Twitter 'garden hose' which offers 10% of tweets for free to approved research projects.

Even if the desired data can be obtained, there are numerous challenges associated with analysing it. Mobile data is qualitatively incomplete. It offers a way to 'see' population displacement, but can only estimate that people are moving, and not why they are moving or what they need. Moreover, the data is only available from places where the mobile network is functional, so there may be gaps and biases that the researcher cannot evaluate. There is also an issue of data source bias. Where the data comes from only one operator, does that skew the research findings toward a particular demographic? The researcher must consider whether the data covers the entire country or territory in question. Many of these same issues apply to social media data, in addition to the challenge of establishing the veracity of messages sent in the highly self-aware context of a social network.

3.2 Facilitating Information Exchange

The swift increase in use of digital communication platforms such as mobile phones and the Internet is also creating opportunities for activism using crowdsourced data in LMICs. Advocates are gaining the ability to aggregate data submitted by individuals more effectively than before, and present it in new ways that can motivate people to action using social media or dedicated platforms. This process is being facilitated by developers acting as intermediaries for advocacy organizations and campaigners. These developers are lowering the barriers of entry through new online platforms that enable advocates to increase the scale of the data they use. Ushahidi is one prominent example of such a platform providing space for collaborative data harvesting (Case Study 3).

Case Study 3: Ushahidi

Ushahidi is a non-profit technology company which has evolved into a platform and advisory service for various users. This evolution has changed their role: now they make the platform, but the people who deploy it have to decide how they connect to the people they are trying to reach. Ushahidi's platform aggregates the lessons of past projects as well as real-time information, making toolkits that combine lessons from many implementations in a similar space.

Their work consists of forums, wikis, face-to-face meetings and meetups. They provide an API with csv and XML download options, and recently released a firehose (crisis.net) that moves data into a similar format for purposes of activism and advocacy.

Ushahidi's current deployments include analyses of environmental data; traffic offence reporting; and a 'Stock out' project to provide information about what sort of medicine is being sent to which health centre, and which releases information about how government disseminates medical supplies to different areas and allows people to ask for more information.

Ushahidi is deployed in numerous contexts, each with different sources and uses of data. In contrast, the work of the Grameen Foundation demonstrates how large datasets can be merged and linked around a specific base layer of mobile phone data (Case Study 4).

Yet with the great potential of information exchange platforms come various difficulties. The problems here differ somewhat from those outlined in the previous section regarding observed data. In this context we are talking mostly about *volunteered* data emerging from multiple – often myriad – different sources. Thus the challenges in this area are as much about aggregating, curating and frequently sharing data as about accessing it. The Grameen AppLab has found that data quality and interpretation can be problematic, for two reasons. First, because sometimes inputs are bad – survey respondents may answer questions incorrectly. Second, the data is generated under locally specific conditions, reflects clients' engagement with specific processes of farming and finance, and this can make the data difficult to analyses remotely without local understanding. This is more pronounced in the case of Ushahidi, as its changing role from direct service provider to (effectively) franchise manager makes it harder to know who is using the platform and how.

Case Study 4: The Grameen Foundation

Grameen Bank's work on microfinance for the poor and unbanked now extends into much of Africa. The related Grameen Foundation has set up an AppLab in Kampala to develop tools which can make use of the information stemming from its clients' microfinance transactions and related information flows. The Bank's target population is those living on less than $2 a day, and working through community knowledge workers (CKWs) it reaches out to clients to form knowledge networks for agricultural information sharing. The AppLab processes and analyses the large datasets stemming from two mobile-based information-sharing applications: 'CKW Search' for information on livestock, crops, and weather; and Pulse messaging – a mobile app which generates popup survey questions and then syncs clients' questions and answers over the web. Through these apps the Grameen Foundation is able to check clients' farms' performance and evaluate its agricultural program.

The apps generate large amounts of data: over four years two million searches have been conducted by microfinance clients, each of which is attached to a base dataset on those clients containing a large number of variables on location, poverty level and individual characteristics. Using these linked datasets and ArcGIS or QuantumGIS to do in-depth analysis, the AppLab can track the quantity and type of requests by region and identify local problems. Successes of the AppLab's work so far include the identification of a new crop pest due to an unusual volume of new searches in a particular region. They sent out an agricultural extension officer to check and the pest was dealt with before it could spread.

There is an important balance to strike in this area between data sharing and privacy, given the sensitivity of some of the data collected, which includes information around clients' income level, expenditure and the size of their family. Uganda has minimal regulations regarding customer privacy in communications networks[1], leaving it up to Grameen Foundation to cooperate with mobile network providers around devising solutions for data protection and privacy challenges.

3.3 Accountability and advocacy

The third main area in which big data is being used to further positive social change is in holding the powerful to account and, as a corollary, advocating for the rights and protections of the less powerful. In the case of accountability, data can be used for numerous functions, such as drawing connections between existing power brokers and elites, and exposing information or the lack of it. One example of this sort of data activism is the Black Monday movement in Uganda (Case Study 5).

Case Study 5: Black Monday

The Black Monday movement was initiated by Ugandan civil society organizations in 2012 to publicize stories of corruption. The activism involved in Black Monday is both locally rooted, publicizing local problems and protests, and global in the technology it uses (websites and social media). the organizers of Black Monday have created an online public space for the debate and criticism of powerful interests, which has gone beyond its original goal to become part of the country's political arena.

[1] https://www.privacyinternational.org/reports/uganda/iii-privacy-issues

Another prominent example of this sort of activism emerges from the work of Chequeado (Case Study 6).

Case Study 6: Chequeado

Chequeado is an Argentine NGO which conducted a factchecking project in recent presidential campaigns and for other political debates. The data involved consisted of crowdsourced references, sources, facts, articles and questions from 40,000 individual participants. Additional outlets include weekly columns in major newspapers and also radio programmes. The project has inspired spinoffs in Colombia, Chile, Costa Rica and also some initiatives in EU, Africa.

The examples of Black Monday and Chequeado show the potential strength of big data when deployed against traditional bastions of control. Yet along with this new power to expose goes the threat of too much exposure. When data concerns the powerless rather than the powerful, the purpose of activism shifts from data exposure to data protection. Thus, another facet of data activism deals with people's own data profiles, and providing the tools for organisations and activists to protect themselves and the people they are advocating for from unwanted digital surveillance or the inappropriate use of their tools and information. This makes generating data awareness and data sovereignty an important activity in parallel with other data activism efforts.

Conducting advocacy and accountability efforts using big data raises some particular challenges since it takes place at the interface between the organizational level, the messy human level where data is produced, and the technical level where it is processed and shaped. Whilst the familiar problems discussed above in terms of handling and interpreting data recur here, there are also more philosophical questions raised by in this domain especially. Through the examples noted here, we can see how data occupies contrasting positions depending on who it relates to. In the context of rulers, data is something to be used offensively, to expose corruption and provide a counterweight to traditional authority, while in the case of the ruled, data should be protected and defended. In other words, in an activist context data is not neutral, but instead imbued with political meaning. As such, the uses to which data is put can be more or less virtuous. Existing unequal power relations and categorizations can thus be reinforced or usurped. This is less a specific challenge and more a general note about how the powerful potentials and dangers of big data in the activist context can intersect.

4. OVERCOMING OBSTACLES

This paper has investigated three motivations for using big data to advance positive social change in LIMCs, through the description and prediction of events (including crises); through the efficient and equitable exchange of information; and through accountability and advocacy as a means of reconfiguring traditional power structures. In each section, we highlighted some of the core challenges specific to each domain. We now conclude with some comments about some of the major challenges raised here and how they can be overcome (see also [2]).

4.1 Access and capacity

Getting access to data and gaining the capacity to handle and manipulate it has repeatedly emerged as a core challenge. The discussions above suggest that the supposed dichotomy of 'open' versus 'closed' data would be better thought of as a scale with 'fully open' and 'fully closed' at either end. Our discussions turned more on 'opening data' – which emphasises the aspects of process and ongoing aspiration – and which can be applied to commercially generated data as much as to public-sector datasets, and data which can lie in between, like mobile phone datasets and financial records. For open data infrastructures to evolve, it is necessary to resolve the gaps that exist and create a broader ecosystem of connected actors, rather than involving only technologists. The situation is similar in the case of capacity-building. The resources to store, analyze, and share large amounts of data could be increased through the use of collaborative learning environments like consortia of universities, and through encouraging open tool development. More experimental funding schemes, including follow-through mechanisms and crowd-funding initiatives could also be used to build capacity within smaller organizations.

4.2 Literacy

Data literacy, or the ability of users to understand the origin, nature and consequences of data, emerged as another core challenge in our discussions. An important aspect of promoting greater data literacy lies in developing more advanced understandings of how data should be used. Different publics – policy makers, activists, individuals – might not require the same level or type of understanding about a given use of data. Thus different strategies can be employed, from building organizational capacity internally to popularizing the negative consequences of data misuse for the public at large. Other remedies might include disclaimers in data journalism and visualization to educate about the misrepresentation of data, and improved teaching of quantitative skills in academic contexts.

4.3 Sensitivity

The third core challenge surrounds issues of data privacy and security. Many of the organizations profiled in this paper handle sensitive data, and strategies are needed for the protection of people whom it concerns. This can be tackled in a number of ways. First, other sectors outside the activism sphere offer models for promoting data security, including institutional review boards in academia and the concept of the 'algorithm ombudsman' [3]. Second, there are existing templates for the evaluation of data sources, e.g. Rebecca McKinnon's Ranking Digital Rights project, and other legal principles which could be appropriated, for example 'do not harm' agreements. Finally, leadership is required within the tech community about how to think critically about existing data standards, including the use of certified courses and ambassadors. Above all, an important first step is for all those involved with data, at each stage of the process, to develop greater awareness as to the sensitivity of what they handle and the negative consequences which can emerge from misuse.

5. CONCLUSION

This position paper has explored uses of big data within an LMIC context, and some of the ways in which sources of big data are currently being used by practitioners, including challenges as well as offering an initial set of solutions and recommendations. It is hoped that the discussions provoked by the meeting at the Rockefeller Foundation will lead over time to a fuller understanding of the potentials and pitfalls of using big data for positive social change in LMIC countries – amongst those who work in this sphere as well as those who are impacted by it.

6. ACKNOWLEDGMENTS

We gratefully acknowledge funding from the Rockefeller Foundation for supporting the conference 'Big data and social change in the developing world' at the Bellagio Centre in Lake Como, Italy, May 12-16 2014. We also acknowledge funding from the Sloan Foundation for the project 'Accessing and Using Big Data to Advance Social Science Knowledge' (2012-2014).

7. REFERENCES

[1] Schroeder, R. 2014. 'Big Data: Towards a More Scientific Social Science and Humanities?', in Graham, M., and Dutton, W. H. (eds.), Society and the Internet. Oxford: Oxford University Press, 164-176.

[2] Schroeder, Ralph, & Taylor, Linnet (Forthcoming). 'Is bigger better? The emergence of big data as a tool for international development policy.' GeoJournal.

[3] Mayer-Schoenberger, Viktor and Cukier, Kenneth. 2013. Big Data: A Revolution that will transform how we live, work and think. London: John Murray.

8. APPENDICES

8.1 Appendix A: Case Study Websites

Case Study 1: Flowminder – www.flowminder.org

Case Study 2: UN Global Pulse – http://post2015.unglobalpulse.net

Case Study 3: Ushahidi – www.ushahidi.com

Case Study 4: Grameen Foundation – www.grameenfoundation.org

Case Study 6: Chequeado – www.cheaqueado.com

8.2 Appendix B: List of Conference Participants

Gilbert Byarugaba Agaba, Grameen Foundation, Uganda

Francis Akindès, Alassane Ouattara University, Côte d'Ivoire

Linus Bengtsson, Flowminder, Sweden

Josh Cowls, Oxford University, UK

Maya Indira Ganesh, Tactical Technology Collective, India

Nimi Hoffman, Rhodes University, South Africa

William Hoffman, World Economic Forum, USA

Anoush Tatevossian, Global Pulse, United Nations

Laura Mann, Leiden Africa Studies Centre, Netherlands

Ulrich Mans, Centre for Innovation, Leiden University, Netherlands

Franziska Meissner, European University Institute, Italy

Eric Meyer, Oxford University, UK

Leonida Mutuku, iHub, Kenya

Sophie Nampewo, Advocates Coalition for Development and Environment (ACODE), Uganda

Angela Oduor, Ushahidi, Kenya

Carly Nyst, Privacy International, UK

Karin Pfeffer, University of Amsterdam, Netherlands

Ralph Schroeder, Oxford University, UK

Nishant Shah, Centre for Internet and Society, India

Pål Sundsøy, Telenor, Norway

Linnet Taylor, University of Amsterdam, Netherlands

Laura Zommer, Chequeado.com, Argentina