data2x

partnering for a
gender data revolution

The Landscape of Big Data for Development

Key Actors and Major Research Themes

Bapu Vaitla
UN Foundation/Data2X

May 2014

data2x | partnering for a gender data revolution

The Landscape of Big Data for Development

Key Actors and Major Research Themes

Bapu Vaitla
UN Foundation/Data2X

Contents

Executive Summary ii
Acknowledgements iii
Abbreviations iii
Introduction: How Big Data Development Research is Different 1
A. Data Exhaust 2
1. Mobile Phone 2
2. Other Types of Data Exhaust 5
B. Online Activity 6
1. Twitter 6
2. Google 8
3. Other Online Activity 9
C. Sensing Technologies 9
1. Remote Sensing 10
2. Personal Sensing 11
D. Crowdsourcing 12
1. Humanitarian Emergencies 12
2. Development 12
E. Non-Research Activities 14
Conclusion 15
References 16

Executive Summary

This report summarizes the current landscape of big data for development[1], considering in particular data exhaust (e.g. cell phone records), online activity (e.g. social media), sensing technologies (e.g. satellite data), and crowdsourced information. It reviews the major big data research initiatives over the past few years and discusses the role of the private sector, academia, multilateral institutions, foundations, donor agencies, and NGOs in these projects. The report relies on a comprehensive review of both the peer-reviewed and grey literature, followed by a series of interviews with researchers working in the big data for development field, each of whom reviewed and provided feedback on the draft.

The use of big data in development is largely being driven by opportunistic partnerships between private companies and academics. Data exhaust is often owned by the private sector, especially mobile phone operators. Online activity, sensing data, and crowdsourced information are often publicly accessible, but the size and complexity of these data sets requires specialized analytical skills. Because of this, and because big data analytics are still in a nascent phase methodologically, academics currently have a high degree of influence in how big data is actually utilized. Some of these academics work in-house for telecommunications and IT firms, but most are in public and private university systems. Multilateral institutions, especially UN Global Pulse, play a key role in publicizing the potential role of big data in development. Foundations fund big data research through a variety of financing streams, and have been important in creating forums where big data researchers can exchange ideas and data sets. The current landscape of big data is, overall, less the result of agenda setting by a small group of politically and economically powerful institutions than it is the unplanned aggregate of diverse projects focusing on those aspects of big data analytics that are methodologically and legally tractable.

Mobile phone call detail records dominate the category of data exhaust, and researchers have focused on how this information can be used to make inferences about human mobility patterns, social network structure, and socioeconomic welfare. Mobility research, for example, has helped to trace transmission patterns of epidemic disease, especially malaria and cholera. Satellite data on environmental and infrastructural variables is the most common type of sensing information, used to map spatiotemporal patterns of illness risk and economic development. Twitter feeds and Google searches are the dominant types of online activity data, and are most commonly used as early warning systems for epidemics and more recently for “sentiment analysis” of users — a snapshot of psychological states, reactions to events, and cultural attitudes. NGOs have generally taken the lead in implementing crowdsourcing applications in humanitarian emergencies, and multilateral institutions have done the same in development.

Contrary to popular perception, big data is not a replacement for traditional data systems. In fact, in the short-term big data projects will need to rely on complementary “ground-truthing” data from traditional sources in order to assess the nature and magnitude of bias in big data sets. Such validation procedures are necessary for end-users of the data, including policymakers, to interpret the contextual meaning of big data across cultures and economies. In addition, big data sets are *not* by virtue of their size exempt from the conventional requirements of good theoretical and statistical practice, including careful problem identification, model construction, and hypothesis testing.

[1] The phrase “big data for development” is used broadly here; the focus is not only on developing world data sets, but also on projects that have the potential to be applied to developing world issues pertaining to human development.

Acknowledgements

The author would like to thank the Data2X team at the United Nations Foundation – Rebecca Furst-Nichols, Emily Courey Pryor, and Mayra Buvinic – for their guidance and support. I would also like to thank the big data scientists and experts who generously gave of their time for interviews and email correspondence, as well as to look over this draft: Vanessa Frías-Martinez, Linus Bengtsson, Aron Culotta, Caroline Buckee, Anoush Rima Tatevossian, Mark Dredze, Michael Paul, Joshua Blumenstock, Patrick Meier, and Rumi Chunara. Their knowledge and that of the Data2X team informs the content of this report, although all errors are solely my own.

Abbreviations

AfDB	African Development Bank
AVHRR	Advanced Very High Resolution Radiometer
CIDA	Canadian International Development Agency
CDR	Call Detail Records
CNES	Centre National d'Études Spatiales (National Center of Space Studies)
DFID	UK's Department for International Development
FIND	Foundation for Innovative New Diagnostics
GPS	Global Positioning System
InSTEDD	Innovative Support to Emergencies, Diseases, and Disasters
MIT	Massachusetts Institute of Technology
MSS	Multispectral Scanner
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NOAA	National Oceanic and Atmospheric Administration
PING	Positive Innovation for the New Generation
QCRI	Qatar Computing Research Institute
SPOT	Système Pour l'Observation de la Terre (Earth Observation System)
TM	Thematic Mapper
UNDP	United Nations Development Program
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development
WFP	World Food Program

Introduction: How Big Data Development Research is Different

This report summarizes the current landscape of big data for development. It reviews the major big data research initiatives over the past few years and discusses the role of the private sector, academia, multilateral institutions, foundations, donor agencies, and non-governmental organizations (NGOs) in these projects.

Following this introduction, each of the subsequent four sections looks at a different type of big data: data exhaust, online activity, sensing technologies, and crowdsourcing[2]. Data exhaust is information passively generated by the use of digital technology, for example cell phones. Online activity encompasses all forms of Internet use, but current research is focused on social media, especially Twitter, and search activity, especially through Google. Sensing technologies conventionally refer to data-gathering satellites and Global Positioning System (GPS)-enabled devices, although a variety of newly invented personal sensors are gathering information on social behavior and environmental conditions. In contrast to the other three types of data, which require little active engagement between user and researcher, crowdsourcing applications actively solicit the knowledge of a wide user base on particular topics or events. The fifth section of the report briefly discusses non-research actors that have been critical in advancing the big data research agenda, especially foundations and donor agencies.

Big data research differs in several aspects from other areas of development. Most of the data used in traditional development research rests with those who have generated it, most commonly government statistical agencies, multilateral databanks, and a few specialized NGOs. Academics generally work with either large-scale data sets controlled by these public and non-profit entities or with smaller data sets they generate themselves specifically for research purposes. Additionally, with the rise of randomized controlled trials and other research designs aimed at feeding more directly into the policy process, partnerships between academics and governments are becoming more common. Data generated by the private sector has been to date a relatively small contributor to development research.

The use of big data, in contrast, is largely being driven by opportunistic partnerships between private companies and academics. Data exhaust is often owned by the private sector, especially mobile phone operators, although legal privacy requirements constrain the ability of the companies to share the data freely. Remote sensing, online activity, and crowdsourced information are often publicly available, but the size and complexity of these data sets require specialized analytical skills. Because of this, and because big data analytics are still in a nascent phase with little methodological consensus, academics currently have a high degree of influence in how big data is actually utilized. Some of these academics work in-house for telecommunications and IT firms, but most are in public and private university systems.

This is not to say that traditional players do not exercise influence. Multilateral institutions like UN Global Pulse have had an important role in marketing the potential of big data for development and initiating research initiatives of their own, in Global Pulse's case leveraging the unique reach of the UN system. The United Nations Children's Fund (UNICEF) in particular has launched several innovative social media and crowdsourcing pilot projects. Private foundations (especially), government donor agencies, and development banks are, as in other areas of development, the primary funding source for big data research initiatives. However, mainly because big data has no objective definitional boundaries, this money is often not earmarked for "big data research" specifically, appearing rather under a range of funding portfolios. After years of skepticism, donor interest in big data for development has increased greatly in the past two years, and for the first time funding for large-scale projects – not simply proofs of concept – appears to be on the horizon.

The current landscape of big data is thus less the result of agenda setting by a small group of politically and economically powerful institutions than it is the unplanned aggregate of diverse projects focusing on those aspects of big data analytics that are methodologically and legally tractable. For example, mobile phone call detail records

[2] Big data does not have a consensus definition. These categories were originally proposed in the UN Global Pulse (2012) report "Big Data for Development: Challenges and Opportunities", written by Emmanuel Letouzé. They are used to narrow the scope of this report to the most relevant current applications of big data to the gender data gap.

dominate the category of data exhaust, and researchers have focused on how this information can be used to make inferences about human mobility patterns and socioeconomic welfare. Satellite data on environmental and infrastructural variables is the most common type of sensing information, used to map spatiotemporal patterns of illness risk and economic development. Twitter feeds and Google searches are the dominant types of online activity data, and they are most commonly used as early warning systems for epidemics and more recently for "sentiment analysis" of users — a snapshot of psychological states, reactions to events, and cultural attitudes. Crowdsourcing is an exception: NGOs and multilateral institutions have driven most crowdsourcing projects.

In sum, in most applications of big data for development the research agenda has been built by individual researchers slowly forming thematic communities, for example around "digital epidemiology" public health research done using cell phone records and social media as the primary data sources. The main challenge in doing big data research well is the same challenge that currently confronts more traditional areas of science: collaboration between disciplines, in the particular case of big data between computer scientists, statisticians, and technical specialists in health, economics, education, and so on. An additional consideration in the development realm, often ignored because big data is seemingly distant from field implementation of programs, is the place-based knowledge that is necessary for not only interpretation but also effective utilization of big data by policymakers and program officers. For example, health surveillance data from online activity may provide a uniquely rich picture of spatial and temporal morbidity trends, but using this information productively depends on local decision-makers and health professionals understanding and trusting the analytical results.

Thus much of the ineffective use of big data – and the (growing) resistance that results from such ineffective use – stems from a lack of sufficient collaboration, manifest especially in the application of innovative computing techniques that unfortunately may not be grounded in contextual knowledge of a place, strong technical understanding of the disciplinary problem, or statistical considerations of accuracy, bias, and precision. In the academic world, this is changing rapidly as collaborative networks strengthen and more big data research enters the peer-reviewed literature and is subject to closer scrutiny; this report profiles some of the best examples of rigorous, policy-relevant big data research for development. Much, however, remains to be done in linking big data research results to real-world policy and program processes.

The sections that follow are not comprehensive, but rather highlight the most well-known projects and major research themes in the categories mentioned above, and the involvement of the various types of development actors in each.

A. Data Exhaust

Data exhaust refers to the "digital footprints" left by behavior — passively generated, digital information related to phone calls, transactions, and the like. At present, data arising from mobile phone use is the most frequently studied form of data exhaust.

1. Mobile Phone

The major source of data exhaust in the developing world is, and will continue to be for the near future, mobile phone Call Detail Records (CDRs) and airtime expense records. Call Detail Records provide data on anonymized caller and receiver phone IDs, the start and end times of calls, call duration, and the location of the caller and receiver (as determined by the location of the closest cell tower). Information on SMS and multimedia content sent is also available. Airtime expense records contain data on the amount of purchase, the time of purchase, existing balance, the phone user's ID, and the nearest tower location at the time of purchase (UN Global Pulse 2013e).

The overall goal is behavioral inference (Eagle 2008). By tracing the social, spatial, and temporal patterns of use, researchers can study a variety of behavioral phenomena, ranging from human and disease mobility patterns, socioeconomic welfare, and the structure of social networks. A few examples of innovative and

well-known projects and researchers are summarized below.

In the private sector, **Telefónica** I+D, based out of Madrid, Spain, has been an important force in creating new approaches to analyzing cell phone data. The company is the research and development wing of the Telefónica Group, one of the largest telecommunications corporations in the world, serving over 200 million users in Latin America, Europe, and the United States (Telefónica 2014). **Vanessa Frías-Martinez** (previously with Telefónica, now with the University of Maryland) and **Enrique Frías-Martinez**, in particular, have been pioneers in the development of algorithms for inferring socioeconomic welfare from mobile phone use patterns, drawing especially on Telefónica's urban data sets in Latin America (V. Frías-Martinez, Virseda, et al. 2010b; V. Frías-Martinez et al. 2013; Soto et al. 2011). Their other work includes the creation of a algorithm for predicting the sex of cell phone users from call detail and airtime purchase records — a critical prerequisite for using these data sets in a gender-disaggregated manner — as well as tools for the creation of census maps and an empirical investigation of how governmental mobility restrictions impacted the spread of the H1N1 virus in Mexico in 2009 (V. Frías-Martinez, E. Frías-Martinez, et al. 2010a; V. Frías-Martinez et al. 2012; E. Frías-Martinez et al. 2011).

In June 2012, the the French telecommunications corporation **Orange Telecom** launched its "Data for Development" initiative, which invited researchers to submit proposals on how anonymous cell phone information could be used for development purposes. The applicants were granted access to a data set of 2.5 billion call records made by five million Orange users in Cote d'Ivôire over a five month period between December 2011 and April 2012 (Blondel et al. 2012). The winners included projects linking social networks and the spread of epidemic disease; mapping mobility patterns through CDRs to help design better public transport networks; and inferring ethnic and other types of social divisions from calling behavior (Berlingerio et al. 2013). This was one of the first instances of a private mobile phone company releasing anonymized records to the public, not only to in-house research departments or individual researchers[3].

There is an expanding network of academic researchers focused on cell phone data. At the **University of Louvain** in Belgium, **Vincent Blondel**'s lab has studied how cell phone data can help produce poverty maps at much greater resolution than typical government statistical surveys

Actors highlighted in **Mobile Phone**

Private Sector

- Telefónica, Enrique Frías-Martinez
- Orange Telecom
- Jana
- Digicel

NGOs

- Flowminder.org

Academia/Research Institutes

- Vanessa Frías-Martinez, University of Maryland
- Nathan Eagle, Harvard University/ Northeastern University
- Engineering Social Systems Lab, Harvard University
- Joshua Blumenstock, University of Washington
- Vincent Blondel, University of Louvain
- Johan Bollen, Indiana University
- Linus Bengtsson, Karolinska Institutet
- Rumi Chunara, Harvard Medical School
- Amy Wesolowski, Carnegie Mellon
- Caroline Buckee, Harvard University
- Xin Lu, Stockholm University
- Erik Wetter, Stockholm School of Economics
- Andy Tatem, University of Southampton
- Petter Holme, Umeå University
- Dan Gillick, University of California-Berkeley

[3] It is also worth noting the vast majority of the in-house research conducted by telecommunications companies is kept out of public view, related as it is to the economic and intellectual property interests of the firms.

[4] "Machine learning", a topic within artificial intelligence, refers to the study and development of computer systems that can recognize and learn from patterns within datasets without explicit step-by-step instruction by programmers.

(Gutierrez et al. 2013). **Johan Bollen**'s group at **Indiana University** has investigated how mobile records illuminate geographical patterns of economic development (Mao et al. 2013).

Nathan Eagle, who holds joint professorships in epidemiology and computer science at **Harvard** and **Northeastern** universities, has looked at how social network structure can be inferred from cell phone data, and how the diversity of contacts in a social network impacts economic development (Eagle, Pentland, et al. 2009b; Eagle et al. 2010). Among numerous other works, Eagle has also used cell phone data to make comparisons between rural and urban communication and mobility behavior (Eagle, de Montjoye, et al. 2009a). One of Eagle's affiliations, the **Engineering Social Systems Lab at Harvard**, is a center of cell phone-based studies, and he is the CEO of **Jana**, a company that offers free phone credit in exchange for mobile subscriber engagement, including participation in research.

Joshua Blumenstock at the **University of Washington** has done similar work to the Telefónica labs, looking at how wealth and migration data can be extracted from cell phone use (Blumenstock et al. 2010; Blumenstock 2012). Blumenstock and Eagle together have also examined gender and class differences in cell phone use, information critical for making accurate inferences from phone data (Blumenstock & Eagle 2012a). Blumenstock and Eagle, along with their colleague **Dan Gillick** at the **University of California-Berkeley**, are among the few researchers to attempt to predict the sex of cell phone users using a machine learning algorithm[4], although their model, using Rwandan CDR data, succeeds only marginally (Blumenstock & Eagle 2012b). Their work, as well as that of Vanessa and Enrique Frías-Martinez, suggests that the utility of such sex-prediction algorithms may vary considerably from context to context. "Ground-truthing" validation surveys using conventional research methods are needed to gauge the context-specific accuracy of models.

Various researchers have analyzed population movements following the 2010 earthquake and subsequent cholera epidemic in Haiti, using data provided by **Digicel**, the country's largest mobile operator. A team lead by **Linus Bengtsson** of the **Karolinksa Institute** in Sweden, in partnership with **Columbia University**, established the position of nearly two million SIM card holders before and after the earthquake, finding that one-fifth of Port-au-Prince's residents left the city by three weeks after the disaster (Bengtsson et al. 2011; Lu et al. 2012). Other applications of CDR data sets to public health are expanding. **Rumi Chunara** at Harvard Medical School and collaborators found that crowdsourced information on cholera cases in Haiti correlated well with government estimates, and was available considerably earlier (Chunara et al. 2012). **Amy Wesolowski** at **Carnegie Mellon**, **Caroline Buckee** at Harvard's School of Public Health, and various collaborators have reconstructed malaria transmission pathways in Kenya from human mobility patterns implied by CDRs (Wesolowski et al. 2012).

Bengtsson is also the director and – with **Xin Lu** of **Stockholm University** and **Erik Wetter** of the **Stockholm School of Economics** – co-founder of **Flowminder.org**, a non-profit organization that analyzes CDRs and remote sensing data for public health. The organization, which counts Caroline Buckee, **Andy Tatem** of the University of Southampton, Joshua Blumenstock, **Petter Holme** of **Umeå University**, and Amy Wesolowski as collaborators, focused initially on looking at the relationship between human mobility patterns and public health outcomes following natural disasters and conflicts, but has increasingly been involved in the analysis of similar issues in development contexts as well.

UN Global Pulse, an initiative of the UN Secretary-General launched in 2009, is an important player in big data for development, piloting a number of innovative applications through partnerships with the private sector and academia. Although Global Pulse has been active in publicizing the utility of call detail records in development research (UN Global Pulse 2013e), most of their own research partnerships deal with other forms of big data, and are accordingly discussed at greater length in other sections. Similarly, various UN agencies have been proactive in utilizing big data, but as with UN Global Pulse, to date the focus has not

been on cell phone records, but rather on online activity analysis and crowdsourcing applications.

2. Other Types of Data Exhaust

Many other types of data exhaust are waiting to be tapped for analysis. One notable example is the data generated by M-PESA, the mobile money application developed through a partnership of **Safaricom**, Kenya's largest mobile network provider, and **Vodafone** (pilot funding for the project came from the **UK's Department for International Development**, DFID). M-PESA allows users to deposit money into mobile phone-based accounts and transfer this money to other account holders. The money in accounts can be retrieved through airtime retailers, who essentially serve as both automated teller machines and wire transfer agents. A related application, M-Shwari, also provides savings and micro-loans services. In 2012, two-thirds of Kenya's adult population had M-PESA accounts and an estimated one-quarter of the country's GDP flowed through the service. Data analysis on transfer, deposit, and withdrawal patterns of M-PESA's 17 million account holders would paint a rich spatially and temporally disaggregated portrait of socio-economic conditions in the country (Mbiti & Weil 2011; Jack & Suri 2011; Mas & Radcliffe 2010). M-PESA is also currently expanding to Afghanistan, South Africa, and India, as Vodafone partners with local banks in each of those countries.

Another source of information is the flood of data generated by mHealth applications (Vital Wave Consulting 2009). Foundations and donor agencies are heavily investing in mHealth; a notable example is the **mHealth Alliance**, which was launched in 2009 by the **UN Foundation**, the **Vodafone Foundation**, and the **Rockefeller Foundation** to foster collaborative solutions to bring mobile-based health solutions to scale in low and middle income countries. A review of the range of mHealth applications is beyond the scope of this report, but of particular interest to the developing world is surveillance and patient monitoring information transmitted by community health workers to central databases. Many of these initiatives remain small-scale or in pilot stage (WHO 2011), but existing data sets could be mined to understand local illness trends, and data formats harmonized to facilitate analysis using big data mining and machine learning techniques. In Thailand and Cambodia, the NGO **Innovative Support to Emergencies, Diseases, and Disasters** (InSTEDD) is working with government departments to facilitate epidemic reporting through cross-platform group chat software (InSTEDD 2006). An SMS illness surveillance system set up by the

Actors highlighted in
Other Types of Data Exhaust

Private Sector

- Safaricom
- Vodafone
- Hewlett-Packard
- Mascom

Donor Agencies and Foundations

- UK Department for International Development (DFID)
- UN Foundation
- Vodafone Foundation
- Rockefeller Foundation
- Clinton Foundation

NGOs

- Innovative Support to Emergencies, Diseases, and Disasters (InSTEDD)
- Positive Innovation for the New Generation (PING)
- Malaria No More UK
- Foundation for Innovative New Diagnostics (FIND)

Government Departments

- Madagascar Ministry of Health
- Botswana Ministry of Health
- Uganda Ministry of Health

Academia/Research Institutes

- Institut Pasteur
- Earth Institute, Columbia University

Multilateral Institutions

- World Health Organization (WHO)
- Innovation Lab, United Nation's Children's Fund (UNICEF)

Madagascar Ministry of Health and **Institut Pasteur** is operational in 13 health centers in the country, and has successfully provided early warnings of multiple disease outbreaks (Randrianasolo et al. 2010). In Botswana, **Hewlett-Packard**, **Botswana's Ministry of Health**, the **Clinton Foundation**, the national mHealth NGO **Positive Innovation for the New Generation** (PING), **Malaria No More UK**, and the cell network provider **Mascom** have partnered to pilot SMS digital surveillance of malaria (PING 2012). Another cell phone-based malaria surveillance project set up by the **World Health Organization**, the **Foundation for Innovative New Diagnostics** (FIND**)**, **Columbia University's Earth Institute**, and the **Ministry of Health in Uganda** yielded impressive results (Asiimwe et al. 2011). Dozens of similar mHealth projects exist (Royal Tropical Institute 2013). Many of these applications are facilitated by open-source and low-cost SMS frameworks like FrontlineSMS and RapidSMS, the latter developed initially by the **UNICEF Innovation Lab** (FrontlineSMS 2013; RapidSMS 2013). It is worth noting that mHealth is one of the few areas in which developing world government agencies, in this case Ministries of Health, are strongly engaging with the kinds of data profiled in this report – a critical prerequisite to big data research of any kind having a significant real world impact.

B. Online Activity

This section describes the landscape of online activity data, an area dominated by information from Twitter feeds and Google searches. Although developing world applications of these types of data are limited, the purposes to which they have been used elsewhere — especially in public health surveillance — are highly pertinent to development efforts. The third subsection below looks at other forms of online activity data.

1. Twitter

The automated analysis of social media is currently one of the most important goals in computational linguistics. Beyond simply counting the number of mentions of a particular topic, "sentiment analysis" and "opinion mining" techniques seek to identify the ideas and emotions expressed in feeds. Such analysis sheds light on the underlying attitudes of social media users that give rise to these ideas and emotions, how these attitudes move through social networks, are opposed or confirmed, and evolve in content (UNICEF et al. 2012; Liu 2012; Pang & Lee 2008).

Early sentiment analysis work attempted to flag comments as merely positive, negative, or neutral. More recent efforts address advanced topics like emotion intensity and analysis across languages (Paltoglou & Thelwall 2012; Tromp 2012). Instead of programmers specifying which words and phrases connote which sentiments, modern natural language processing (NLP) methods rely on machine learning algorithms that apply lessons learned from the examination of "training" data sets to the interpretation of new data sets. Sentiment analysis of social media faces many challenges, including unconventional language formulations, the use of slang, and missing information about context (Bifet &

Actors highlighted in **Twitter**

Private Sector

- Twitter
- Crimson Hexagon
- DataSift
- Microsoft Research

Donor Agencies and Foundations

- UN Foundation
- Qatar Foundation: Qatar Computing Research Institute (QCRI)

Academia/Research Institutes

- Aron Culotta, Illinois Institute of Technology
- Michael J. Paul and Mark Dredze, Johns Hopkins University
- Johan Bollen, Indiana University
- Munmun De Choudhury, Georgia Tech University
- Data-Pop Alliance, Harvard Humanitarian Initiative, MIT Media Lab, and Overseas Development Institute

Multilateral Institutions

- UN Global Pulse
- UN Millennium Campaign
- UNICEF

Frank 2010; Power et al. 2010). Machine learning methods, however, are increasingly able to handle such heterogeneity (Barbosa & Feng 2010; Collier et al. 2011)

At this point, **Twitter** is the main source of sentiment and opinion data, due to its public accessibility, the amount of information available, user diversity, and the range of topics discussed (Pak & Paroubek 2010). **UN Global Pulse** is carrying out several Twitter-centered projects. One, in conjunction with the **UN Foundation**, looks at whether the Every Woman Every Child program has been effective in raising awareness of child and women's health issues, as measured by number of tweets (UN Global Pulse 2013a). A Twitter sentiment analysis performed by **Crimson Hexagon** and UN Global Pulse in Indonesia finds that conversation about economic concerns revolves around four subjects, "housing, gas/fuel, personal finance, and food" (Lopez & St. Amand 2013). The monitoring system examines weekly and monthly patterns in trending topics, and serves as a surveillance system for tracking prices of key staple foods like rice (UN Global Pulse 2013f). The project finds that Twitter feeds pick up short-term concerns better than longer-term goals (Crimson Hexagon & UN Global Pulse 2011). UN Global Pulse also has an ongoing partnership with the **UN Millennium Campaign** and **DataSift** to scan Twitter for the most commonly discussed development-related topics worldwide. The results suggest that, in contrast to surveys that ask about long-term priorities, analysis of Twitter feeds enables a unique understanding of "daily hopes and grievances", although the data is certainly biased towards the young and better-off (Nielsen 2013). Among UN bodies, **UNICEF** recently undertook a study of social media content in Eastern Europe to look at attitudes towards vaccination (Majewski & Beger 2013). A project of the **Qatar Foundation**-funded **Qatar Computing Research Institute (QCRI)** and the **Data-Pop Alliance**, a joint initiative of the Harvard Humanitarian Initiative, the MIT Media Lab, and the Overseas Development Institute, is looking at whether Twitter sentiments can indicate trends of poverty and unemployment (Meier 2013). More broadly, the Data-Pop Alliance is engaged in a diverse body of work around big data and development, including compiling a comprehensive e-library of past and ongoing research, working with national statistical offices to incorporate big data, and conducting primary research on key methodological issues around accurate and rigorous analysis of big data. They are also a key actor in kick-starting the dialogue around the political economy of big data, including confronting the issue of big data's impact on the democratization of information for development.

Epidemic surveillance through Twitter feeds has been a major focus of academic research (Krieck & Dreesman 2011; Lopes et al. 2009). Influenza is by far the most studied illness. To provide just one example, **Aron Culotta** of the **Illinois Institute of Technology**, using only a small set of matching keywords, finds a 95% correlation between a set of 570 million tweets in 2009-10 and official health statistics (Culotta 2010). The list of researchers and universities involved with influenza surveillance through Twitter is too long to include here, but some of the the key studies include those done by Aramaki et al. (2011), Lampos & Cristianini (2010), and Achrekar et al. (2011). **Michael J. Paul** and **Mark Dredze** at **Johns Hopkins University** have done similar work looking at illnesses other than influenza, and identify behavioral risk factors, symptoms, and treatments used (Paul & Dredze 2011a; Paul & Dredze 2011b). Twitter feed content can also be combined with user locations and the interaction of that user with friends on social media to predict future health states — that is, to trace the spread of an illness through social networks (Sadilek et al. 2012). Illness spreading can also be tracked geographically if Twitter users elect to geo-tag their tweets.

Twitter applications are diversifying. **Munmun De Choudhury** (now at **Georgia Tech University**) and her collaborators at **Microsoft Research** have looked at links between clinically diagnosed depression and social media expression, and used these correlations to identify signals of depression among Twitter users who have not been clinically diagnosed (De Choudhury et al. 2013a). De Choudhury's team has

also used Twitter data to predict emotional states among postpartum mothers (De Choudhury et al. 2013b). **Johan Bollen** at **Indiana University** and his collaborators looked at all tweets in a five-month period at the beginning of the global recession in 2008. They use a psychological tool called the Profile of Mood States to test the extent to which economic hardship led to emotions like tension, depression, anger, fatigue, and confusion, as picked up by Twitter feeds (Bollen et al. 2009). Fluctuations measured by similar Twitter-based mood tracking tools can be predictive of economic trends at the population level, for example changes in stock market averages (Bollen et al. 2010). Twitter feeds have also been analyzed as an inexpensive means of gauging political views (Stieglitz & Dang-Xuan 2012; Tumasjan et al. 2010). Correlations of tweets with polls and other more commonly used means of gauging political sentiment tend to vary across contexts, but improvements in textual analysis and identification of bias are likely to improve the performance of current methods (O'Connor et al. 2010). Twitter itself recently launched a data grant program to allow selected researchers access to data sets (Krikorian 2014).

2. Google

Google search data has been useful in epidemic surveillance. **Google Flu Trends**, building on the pioneering work of **Gunther Eysenbach** (2006) of the University of Toronto and others, and operated by Google itself, is currently the most widely studied application (Google 2014a). The service works by correlating tens of millions of search queries with official data on influenza-like illness from the Centers for Disease Control (CDC) (Watts 2009). The search terms that prove to be the strongest determinants of flu are combined in a linear model to predict incidence in advance (Ginsberg et al. 2008). Flu Trends has been shown to pick up patterns 7 to 10 days before CDC data (Carneiro & Mylonakis 2009), and correlations hold in both the US and Europe (Valdivia et al. 2010). Updated models also show correlation with non-seasonal outbreaks like the H1N1 epidemic in 2009 (Cook et al. 2011). Flu Trends also correlates with "harder" forms of health system data such as positive influenza tests and emergency room intakes of people with flulike symptoms (Dugas et al. 2012 ;Ortiz et al. 2011). Because Google search data is publicly available, external researchers have been able to develop ever stronger algorithms for interpreting illness trend data, controlling for such factors as the degree of Internet activity and the size of a population (Pevaiz et al. 2013).

Actors highlighted in **Google**

Private Sector

- Google

Academia

- Gunther Eysenbach, University of Toronto

Google searches have also been used to track seasonal illnesses other than influenza. Google searches for dengue, for example, correlate extremely well with official statistics on infection across many countries (Althouse et al. 2011). Searches pertaining to Lyme disease also corroborated spatial and temporal patterns in the US (Seifter et al. 2010). Google searches also correlate with certain chronic noninfectious diseases that vary seasonally, including hypertension, heart attacks, diabetes, and kidney stones (Breyer & Eisenberg 2013). **John Ayers** and his collaborators have found that Google searches can also function well for mental health surveillance, with clear distinct seasonal patterns appearing in the United States and Australia for a wide variety of mental health problems (Ayers et al. 2013).

Research on non-health topics (that are relevant to development) using Google search queries is still rare, although there has been some investigation of how searches predict market activity and socioeconomic welfare (Wu 2009; Preis et al. 2012). The creation of **Google Trends**, a site provided by Google to track frequency of search terms disaggregated by country and time period, should facilitate such work in the future (Google 2014b).

3. Other Online Activity

There are many other types of online activity analysis. Twitter, of course, is not the only source of online sentiment information. A partnership between **UN Global Pulse** and the statistical software company **SAS** examined conversations in blogs and forums about jobs and unemployment in Ireland and the US. By classifying the moods revealed in these conversations, the project was able to identify leading, in-process, and trailing indicators of unemployment (UN Global Pulse 2013g; SAS & UN Global Pulse 2011). UN Global Pulse is also partnering with the International Labor Organization and Indonesia's Ministry of National Development Planning to analyze online articles, blogs, and social media to understand women's workplace conditions and public views on women's employment in Indonesia (UN Global Pulse 2014). Although sentiment analysis of **Facebook** posts is complicated by non-public posting and the length of messages, some initial attempts are underway, including a measure of Facebook-based "gross national happiness" in the United States (Kramer 2010; Shrivastava & Pant 2012).

Actors highlighted in **Other Online Activity**

Private Sector

- SAS
- Facebook
- Recorded Future
- PriceStats

NGOs

- DataKind

Academia/Research Institutes

- John Brownstein and Clark Freifeld, Boston's Children Hospital
- Massachusetts Institute of Technology (MIT)

Multilateral Institutions

- UN Global Pulse
- UNDP

Health surveillance through website content analysis is also expanding. **John Brownstein** and **Clark Freifeld** of **Boston Children's Hospital** created HealthMap in 2006, an application to mine reports of epidemic outbreaks from a variety of web sources, including news sites, RSS feeds, and official alerts from health agencies (Brownstein et al. 2009; Brownstein & Freifeld 2007; Brownstein et al. 2008; Freifeld et al. 2008). Various other "digital epidemiology" applications using online activity data are operational or in development (Salathé et al. 2012). More generally, advances in computing power make mining of massive amounts of data — even originally non-digital data, for example from news headlines archives — possible, and this holds great potential for tracking social trends (Leetaru 2011). The **UNDP** worked with **Recorded Future** to scan tens of thousands of online media sources to assess political risk in Georgia prior to the country's 2008 elections (Radojevic 2012).

Price scraping is another form of online activity monitoring. A partnership of UN Global Pulse, **PriceStats**, and the **Massachusetts Institute of Technology**'s Billion Prices Project searches websites for product and price information, noting the source and time of the scraped data. In the pilot project, bread prices were collected between 2007 and 2011 in supermarkets in six South American countries. The data was used to create an "e-Bread index" to estimate inflation in these countries (Rigobon 2011; UN Global Pulse 2013c). Similarly, the data scientist volunteer NGO **DataKind** has developed tools to scrape price websites to estimate inflation in Kenya (DataKind 2013).

C. Sensing Technologies

Although the phrase "big data" is relatively new, the existence of large, complexly structured data sets is not. Of particular note is the enormous amount of information collected by satellites and other remote sensing devices over the past several decades. With these launch of new satellites by middle-income countries over the last few years, the quantity of data has recently expanded even more. In addition, increases in the computing power available to analyze these massive data sets, and new methodologies developed to take advantage of

this computer power, open up exciting possibilities for research. The following sections briefly review the use of remote sensing technologies in development, as well as take a brief look at the new generation of personal "non-remote" passive sensing technologies.

1. Remote Sensing

The most important application of remote sensing to development comes in the mapping of environmental phenomena and human infrastructure, including vegetation, water bodies, transport networks, and land use patterns. Sensing research on spatiotemporal variation in public health conditions, for example malaria vector abundance, dates back more than two decades (see (Beck et al. 1994; Beck et al. 1997), among many others). However, it is only in the last decade or so that this research has been utilized in the policy process through the creation of epidemic early warning systems and other initiatives (Ceccato et al. 2005; Thomson & Connor 2001); as noted above, increased computing power is partially responsible for the more rapid turnaround between data collection and release of analytical results.

Actors highlighted in **Remote Sensing**

Government Departments

- US National Aeronautics and Space Administration (NASA)
- US National Oceanic and Atmospheric Administration (NOAA),
- France's National Center of Space Studies (CNES)

Remote sensing systems can predict the degree of both inter-year and intra-year risk of illness, and thereby help mobilize preventative and curative resources well in advance of actual morbidity and mortality. Although malaria has been by far the most studied disease through remote sensing, the epidemiological landscapes of a wide variety of other illnesses have also been analyzed. An incomplete list includes lyme disease, cholera, meningitis, dengue, Rift Valley fever, schistosomiasis, West Nile fever, and even obesity (Kelly et al. 2011; Beck et al. 2000; Molesworth et al. 2003).

Another important application is food security forecasting based on satellite sensing of vegetation density. The Normalized Difference Vegetation Index (NDVI), the most commonly used measure, is calculated relies on the observation that growing plants absorb radiation in the visible range of the electromagnetic spectrum and reflect light in the near-infrared range (Ceccato et al. 2005). Satellites that gather data across the spectrum are able to thus estimate vegetation density and predict crop yields well before the actual harvest. In combination with household-level information on factors like family size, available labor, input and output costs, and so on, NDVI can help evaluate the risk of impending food insecurity (Quinn et al. 2010). NDVI can also be used be used to predict occurrence of seasonal diseases, which are often determined by changes vegetation density and thus vector habitat (Rahman et al. 2006; Adimi et al. 2010; Machault et al. 2011).

Many other applications of remote sensing data exist. Access to schools, health facilities, and water points is often determined in the developing world by distance, physical obstacles, and transport networks, especially for rural communities. Remote sensing data can help predict seasonal fluctuations in time and transport costs of accessing these resources — as, for example, rainfall affects road quality or river levels — and evaluate these changes in light of intra-year variations in income availability and expenditures. Other ideas are being developed; for example, recent work looks at remote sensing of lighting as an indicator of poverty levels (Noor et al. 2008; The World Bank & DataKind 2013).

One key characteristic of remote sensors involves the tradeoff between spatial and temporal resolution (Hay et al. 2012). For example, accurate evaluation of disease risk requires the use of both high spatial resolution sensors for measurement of factors like land cover and land use and high temporal resolution

sensors to pick up changes in temperature, rainfall, and crop density (Machault et al. 2011). Advances in image analysis of high spatial resolution data are increasingly allowing researchers to identify micro-level variations in, for example, vector habitat (Kelly et al. 2011), which is critically important for assessing disease risk and transmission dynamics in highly heterogeneous environments like urban areas (Tatem & Hay 2004).

Government agencies are the most important sources of freely available remote sensing data. Until recently, most health remote sensing research used data from the United States' **National Aeronautics and Space Administration** (NASA) Landsat Multispectral Scanner (MSS) and Thematic Mapper (TM) satellites, the Advanced Very High Resolution Radiometer (AVHRR) of the **National Oceanic and Atmospheric Administration** (NOAA), and the **French National Center of Space Studies'** (CNES) Earth Observation System (SPOT) (Beck et al. 2000). Many more earth observation satellites have been launched in the past few years by governments around the world (UCS 2013).

2. Personal Sensing

A new generation of personal sensors, embedded in phones, motor vehicles, and other technologies, is being used to passively measure and interpret behavioral signals — a field known as "reality mining" (Eagle & Sandy Pentland 2005). **Alex "Sandy" Pentland's Human Dynamics Lab at the Massachusetts Institute of Technology** is among the leaders in this field. Voice analysis software and accelerometers built into cell phones may be able to diagnose signs of depression and stress. GPS-enabled devices can infer social network structure by recording user movement patterns, and can also be used to locate key public services more efficiently based on the study of the aggregated travel routes of tens of thousands of people (Pentland et al. 2013; Pentland et al. 2009). Bicycle-mounted sensors, such as those developed by **Dartmouth University's Metrosense** project for its Bikenet application, can gather pollution data in cities and develop a spatially and temporally rich portrait of threats to human health. The **University of California-Los Angeles'** (UCLA) **Center for Embedded Network Sensing**, **Intel's Urban Atmospheres initiative**, **MIT**'s **Cartel project**, and the **CitySense** partnership of **Harvard University**, the **City of Cambridge**, and **BBN Technologies** are also important endeavors in the building of personal sensing networks (Campbell et al. 2008).

Actors highlighted in **Personal Sensing**

Private Sector

- BBN Technologies, Raytheon

Academia/Research Institutes

- Alex "Sandy" Pentland and Human Dynamics Lab, MIT
- Metrosense, Dartmouth University
- Center for Embedded Network Sensing, UCLA
- Cartel project, MIT
- Citysense, Harvard University

Government Departments

- City of Cambridge, Massachusetts

To this point, few personal sensing technologies have been applied for development, but the potential is great, particularly in the field of "participatory sensing", wherein individuals and communities are able to choose priorities and carry out projects. As Burke et al. (2006) write,

> *With the right tools, professionals and community groups alike could employ participatory sensing campaigns to gather data about short-term concerns…without waiting for a formal project or grant funding— yielding bottom-up, grassroots sensing. Citizens have intimate knowledge of patterns and anomalies in their communities and enabling them to respond is both empowering and valuable to long-term research…*

The point about "patterns and anomalies" cannot be overemphasized. The overall objective of sensing is to detect such variation, and local knowledge can greatly help researchers design more relevant technologies from the start. This is where the boundary between sensing and crowdsourcing (discussed in the next section) blurs: the creation of a body of local knowledge that is assisted by imports of ideas and technologies, but driven primarily by the wishes and needs of those who stand to benefit most from that knowledge.

D. Crowdsourcing

This section reviews crowdsourcing applications used in both humanitarian and development contexts. NGOs have generally taken the lead in implementing crowdsourcing applications in humanitarian emergencies, and multilateral institutions have done the same in development.

1. Humanitarian Emergencies

The NGO **Ushahidi** first gained notice in 2008 for its efforts in crisis mapping violence in the aftermath of Kenya's elections (Meier 2008). Ushahidi was also praised for its work in helping the United States military and other emergency responders find individuals in need, mainly using **Twitter** reports, following the 2010 Haitian earthquake (Munro 2010). In the early days following the disaster, Ushahidi was in fact the only source of aggregate geospatial information, processing at least 40,000 reports and mapping nearly 4000 individual events (Morrow et al. 2011). Integration with translation services like **Mission 4636** and open source mapping platforms like **OpenStreetMap** helped improve the usability of the data stream.

The Ushahidi experience in Haiti led to a rapid expansion in crisis mapping initiatives. The **International Network of Crisis Mappers**, co-founded by **Patrick Meier** and **Jen Ziemke** in 2009, has grown to include over 6000 members worldwide. The network, supported by a wide range of multilateral organizations and government agencies, universities, and private companies, organizes an annual conference that is an important forum for the exchange of ideas around humanitarian technology. The **Standby Task Force**, created at the 2010 conference, mobilizes volunteers to assist on-the-ground response teams in mapping needs, although its data sources are broader than crowdsourced reports alone. **Sahana** and **Humanity Road** are just two other prominent organizations of the many who use crowdsourcing data as part of their approach to assist in increasing information flow in humanitarian emergencies. Crisis mapping is becoming ever more sophisticated. The current Syria Tracker Crisis Map efforts, built on Ushahidi's Crowdmap platform, combines crowdsourced reports from Twitter, Facebook, and other sources with data mining programs that scour the Web for news of important events, especially killings and human rights violations (Meier 2012).

Actors highlighted in Humanitarian Emergencies

NGOs

- Ushahidi
- Mission 4636
- OpenStreetMap
- International Network of Crisis Mappers: Patrick Meier, Jen Ziemke
- Standby Task Force
- Sahana
- Humanity Road

Private Sector

- Twitter

2. Development

UN Global Pulse is conducting several crowdsourcing projects in non-humanitarian contexts. One massive effort is the partnership between UN Global Pulse and **Jana**, a company that provides free airtime in exchange for user input, to paint a global portrait of human welfare through SMS surveys. The initial survey received over 90,000 responses from more than thirty countries (UN Global Pulse 2013d). Global Pulse

is also partnering with **Question Box**, a service allowing users to ask questions to an Internet search team about any subject, to track temporal trends in people's concerns (UN Global Pulse 2013h). UN Global Pulse's Uganda lab — one of three offices piloting techniques on the country level — is also experimenting with using crowdsourcing within an early warning system for economic and environmental shocks. The idea would be to detect anomalies in data exhaust, such as decreased mobile phone airtime purchases, remittances, or banking activity, and then solicit crowdsourced information from ground-level sources. If such information indicates an incipient crisis, more formal surveys would be launched (Kirkpatrick 2010).

UNICEF's Uganda office is implementing U-Report, a mobile phone application that sends questions on development topics to over 250,000 members (UNICEF Uganda 2014). These "reporters" respond, at no cost, by selecting an option from a pre-formulated menu or sending more detailed replies. The results can be easily disaggregated by sex and age. Although there is a clear bias towards youth in its membership, the application has been useful in capturing general trends of public sentiment and helping to start informational campaigns about key topics. In one instance, an outbreak of nodding disease was identified through U-Report and treatment information sent to the affected area. Automated text classification tools have allowed more detailed analysis of the information arriving from U-Report (Melville et al. 2013). UNICEF Uganda is also experimenting with a range of other digital reporting applications, including a school monitoring, birth registration, disease surveillance, and information on access to health services (Cummins & Huddleston 2013).

Actors highlighted in **Development**

Multilateral Institutions

- UN Global Pulse
- UNICEF Uganda
- World Food Program (WFP)

Private Sector

- Jana
- Question Box
- Mobile Accord

Donor Agencies and Foundations

- USAID
- UK DFID
- Canadian International Development Agency (CIDA)
- Swedish Ministry for Foreign Affairs
- African Development Bank (AfDB)

NGOs

- mWater
- Question Box

Crowdsourcing is also gaining favor among donors. The **United States Agency for International Development** (USAID) provided funding to the nonprofit organization **mWater** to develop a mobile phone application in Tanzania that helps citizens perform water quality tests and upload this information to a database that maps water sources

(mWater 2013). The **World Food Program** (WFP), supported by the Humanitarian Innovation Fund — a joint venture financed by the **UK's DFID**, the **Canadian International Development Agency** (CIDA), and the **Swedish Ministry of Foreign Affairs** — has also launched the mVAM pilot project to collect food security information through SMS surveys in the Democratic Republic of the Congo and Somalia (Humanitarian Innovation Fund 2013; WFP 2013). The **African Development Bank** (AfDB), in partnership with the mobile platform services company **Mobile Accord**, is also using SMS surveys to gauge the impact of AfDB-supported projects (AfDB 2012).

E. Non-Research Activities

This section reviews non-research activities — especially the setting of funding priorities, the publicizing of big data's potential for development, and the creation of forums to exchange ideas and data sets — that have an important impact on the big data research agenda.

As discussed in previous sections, **UN Global Pulse** plays a key role in disseminating information about the applications of big data to development (UN Global Pulse 2013b). In addition to its research partnerships with various private sector companies and UN agencies, Global Pulse has also expanded awareness of big data's potential at a number of conferences.

The **Gates Foundation** is increasingly engaging with big data for development. The last round of the foundation's "Grand Challenges Initiative" awarded six $100,000 grants under the topic of "increasing the interoperability of social good data". The winning projects included ideas to improve information management following natural disasters, incorporating community data on education, developing tools to build cross-data set compatibility, and mapping key community assets like water points and health clinics through a micro-blogging application (Information Week 2013). The Gates Foundation has also invested heavily in big data systems to collect geographical, medical supply, and health service tracking information in its polio eradication program (Greenberg 2014).

Actors highlighted in **Non-Research Activities**

Multilateral Institutions

- UN Global Pulse
- UNDP
- World Bank

Donor Agencies and Foundations

- Gates Foundation
- Ford Foundation
- Skoll Foundation
- Moore Foundation
- Sloan Foundation
- Qatar Foundation: QCRI

NGOs

- DataKind

Other foundations have played important roles. The **Ford Foundation** and the **Skoll Foundation** sponsor two major events around big data, the Wired for Change Conference and the Skoll World Forum (Ford Foundation 2012; Skoll Foundation 2013). The **Moore Foundation** and the **Sloan Foundation** recently launched a 5-year, $38 million initiative to support big data research at New York University, the University of California-Berkeley, and the University of Washington. The **Qatar Computing Research Institute** consults on various projects involving Twitter and cell phone data, led by its director, **Patrick Meier**, formerly the head of crisis mapping at Ushahidi.

Several "DataDives", where organizations and individuals explore data sets and exchange ideas on how to use big data for development, have been held in the last few years. Building on the efforts of **DataKind** — the volunteer data scientist NGO that developed the original "DataDive" concept — the **UNDP** and the **World Bank** held a DataDive in February 2013 in Vienna. With UN Global Pulse and QCRI, the Bank organized a similar "Big Data Exploration" in Washington DC in March 2013. Both events focused on poverty measurement and systems to monitor project-level corruption (Center for Public Administration Research & Open Knowledge Foundation 2013; World Bank & DataKind 2013).

Overall, following a long period of initial skepticism about the use of big data in development, donors have greatly increased their engagement with big data in the past two years. There is currently a window of opportunity for new ideas: funding is increasing and the state of research is democratic and innovative, driven often by unconventional partnerships across academic disciplines and among actors from the private, public, and non-profit sectors.

Conclusion

This report has reviewed the major themes in big data research for development and highlighted the role of key actors. Private companies, especially mobile phone network operators, Google, and Twitter, control much of the data exhaust and online activity data being currently used. The majority of academics working with big data for development focus on behavioral inference from cell phone records, epidemiological surveillance from Google searches, and sentiment analysis from Twitter feeds, although many other smaller strands of big data research are also being pursued. Multilateral institutions, especially UN Global Pulse, play a key role in publicizing the potential of big data in development. Foundations fund big data research through a variety of financing streams, and have been important in creating forums where big data researchers can exchange ideas and data sets. To this point, donors and developing world governments have been focused mostly on smaller crowdsourcing initiatives, as well as opening up data access and creating more user-friendly data management systems, although their scope of interest is expanding. NGOs have thus far been largely limited to data sharing, except in the category of humanitarian crowdsourcing, where tech-savvy organizations have taken the lead.

In closing, one common misperception about the future of big data is worth addressing: the notion that big data can wholly replace more traditional data systems. On the contrary, for the foreseeable future the utility of big data will depend on the creative combining of big data and traditional data sets to analyze development phenomena. For example, mobile phone and social media data sets are not necessarily representative of the behavior of the entire population; in particular, sub-groups with limited access to technology may be under-represented. In other words, big data sets are *not* by virtue of their size exempt from the conventional requirements of good statistical methodology. Ground-truthing research using conventional survey methods is often needed to validate the representativeness of big data, or to identify the nature and magnitude of biases within big data sets. Once these biases are known, the potential of big data – namely, information that is highly spatially disaggregated and is generated frequently, occasionally in real-time – can be exploited in appropriate ways.

References

Achrekar, H., Ghande, A., Lazarus, R., Yu, S., & Liu, B., 2011. Predicting Flu Trends Using Twitter Data. *IEEE INFOCOM 2011 - IEEE Conference on Computer Communications Workshops,* pp. 702–707.

Adimi, F., Soebiyanto, R.P., Safi, N., & Kiang, R., 2010. Research Towards Malaria Risk Prediction in Afghanistan Using Remote Sensing. *Malaria Journal,* 9:125.

Althouse, B.M., Ng, Y.Y. & Cummings, D.A.T., 2011. Prediction of Dengue Incidence Using Search Query Surveillance. *PLoS Neglected Tropical Diseases*, 5(8), p. e1258.

African Development Bank, 2012. *AfDB Goes Mobile for Research Data with Partner Mobile Accord.* Available at: http://www.afdb.org/en/news-and-events/article/afdb-goes-mobile-for-research-data-with-partner-mobile-accord-9354/ [Accessed January 2014].

Aramaki, E., Maskawa, S. & Morita, M., 2011. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* pp.1568–1576.

Asiimwe, C., Gelvin, D., Lee, E., Ben Amor, Y., Quinto, E., Katureebe, C., Sundaram, L., Bell, D., & Berg, D., 2011. Use of an Innovative, Affordable, and Open-Source Short Message Service-Based Tool to Monitor Malaria in Remote Areas of Uganda. *American Journal of Tropical Medicine and Hygiene*, 85(1), pp.26–33.

Ayers, J. W., Althouse, B.M., Allem, J., Rosenquist, J.N., & Ford, D.E., 2013. Seasonality in Seeking Mental Health Information on Google. *American Journal of Preventative Medicine,* 44(5), pp.520-525.

Barbosa, L. & Feng, J., 2010. Robust Sentiment Detection on Twitter From Biased and Noisy Data. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*, pp.36–44.

Beck, L.R., Rodriguez, M.H., Dister, S.W., Rodriguez, A.D., Rejmankova, E., Ulloa, A., Meza, R.A., Roberts, D., Paris, J.F., Spanner, M.A., Washino, R.K., Hacker, C., & Legters, L.J., 1994. Remote Sensing as a Landscape Epidemiologic Tool to Identify Villages at High Risk for Malaria Transmission. *The American Journal of Tropical Medicine and Hygiene,* 51(3), pp.271–280.

Beck, L.R., Rodriguez, M.H., Dister, S.W., Rodriguez, A.D., Washino, R.K., Roberts, D., & Spanner, M.A., 1997. Assessment of a Remote Sensing-Based Model for Predicting Malaria Transmission Risk in Villages of Chiapas, Mexico. *The American Journal of Tropical Medicine and Hygiene,* 56(1), pp.99–106.

Beck, L.R., Lobitz, B.M. & Wood, B.L., 2000. Remote Sensing and Human Health: New Sensors and New Opportunities. *Emerging Infectious Diseases,* 6(3), p.217.

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), p.e1001083.

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 Discovery in Databases*. Springer Berlin Heidelberg, pp. 663–666.

Bifet, A. & Frank, E., 2010. Sentiment Knowledge Discovery in Twitter Streaming Data. *Proceedings of the 13th International Conference on Discovery Science,* pp.1–15.

Blondel, V.D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., & Ziemlicki, C., 2012. Data for Development: the D4D Challenge on Mobile Phone Data. *arXiv preprint arXiv:1210.0137.*

Blumenstock, J.E., 2012. Inferring Patterns of Internal Migration From Mobile Phone Call Records: Evidence From Rwanda. *Information Technology for Development,* 18(2), pp.107–125.

Blumenstock, J.E. & Eagle, N., 2012a. Divided We Call: Disparities in Access and Use of Mobile Phones in Rwanda. *Information Technologies & International Development,* 8(2), pp.1–16.

Blumenstock, J.E., Gillick, D., & Eagle, N., 2012b. Who's Calling? Demographics of Mobile Phone Use in Rwanda. *Association for the Advancement of Artificial Intelligence Spring Symposium,* pp.116-117.

Blumenstock, J.E., Shen, Y. & Eagle, N., 2010. A Method for Estimating the Relationship Between Phone Use and Wealth. *Proceedings of the QualMeetsQuant Workshop at ICTD*.

Bollen, J., Pepe, A. & Mao, H., 2009. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. *ICWSM11* - poster.

Bollen, J., Mao, H., & Zeng, M.L., 2011. Twitter Mood Predicts the Stock Market. *Journal of Computational Science* 2(1), pp.1-8.

Breyer, B.N. & Eisenberg, M.L., 2013. Use of Google in Study of Noninfectious Medical Conditions. *Epidemiology,* 21(3), pp.584–585.

Brownstein, J.S. & Freifeld, C.C., 2007. HealthMap: the Development of Automated Real-Time Internet Surveillance for Epidemic Intelligence. *Eurosurveillance,* 12(48), pp.1–4.

Brownstein, J.S., Freifeld, C.C., Reis, B.Y., & Mandl, K.D., 2008. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine,* 5(7), p.e151.

Brownstein, J.S., Freifeld, C.C. & Madoff, L.C., 2009. Digital Disease Detection—Harnessing the Web for Public Health Surveillance. *New England Journal of Medicine,* 360(21), pp.2153–2157.

Burke, J.A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., & Srivastava, M.B., 2006. Participatory sensing. *UCLA Center for Embedded Network Sensing Papers*.

Campbell, A.T., Lane, N.D., Miluzzo, E., Peterson, R.A., Lu, H., Zheng, X., Musolesi, M., Fodor, K., Eisenman, S.B., & Ahn, G. 2008. The Rise of People-Centric Sensing. *IEEE Internet Computing*, 12(4), pp.12–21.

Carneiro, H.A. & Mylonakis, E., 2009. Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, 49, pp.1557–1564.

Ceccato, P,, Connor, S.J., Jeanne, I., & Thomson, M.C., 2005. Application of Geographical Information Systems and Remote Sensing Technologies for Assessing and Monitoring Malaria Risk. *Parassitologia*, 47(1), pp.81–96.

Center for Public Administration Research & Open Knowledge Foundation, 2013. *Vienna2013*, (Austria). Available at: http://wiki.opendataday.org/Vienna2013 [Accessed January 2014].

Chunara, R., Andrews, J.R. & Brownstein, J.S., 2012. Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *American Journal of Tropical Medicine and Hygiene*, 86(1), pp.39–45.

Collier, N., Son, N.T. & Nguyen, N.M., 2011. OMG U Got Flu? Analysis of Shared Health Messages for Bio-Surveillance. *Journal of Biomedical Semantics*, 2(Suppl 5), p.S9.

Cook, S., Conrad, C., Fowlkes, A.L, & Mohebbi, M.H., 2011. Assessing Google Flu Trends Performance in the United States During the 2009 Influenza Virus a (H1N1) Pandemic. *PloS One,* 6(8), p.e23610.

Crimson Hexagon & UN Global Pulse, 2011. *Twitter and Perceptions of Crisis Related Stress: Methodological White Paper*, Available at: http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress.

Culotta, A., 2010. Towards detecting influenza epidemics by analyzing Twitter messages. pp.115–122.

Cummins, M. & Huddleston, B., 2013. Real Time Monitoring for the Most Vulnerable: UNICEF's Experience in Uganda. *IDS Bulletin,* 44(2), pp.57–68.

DataKind, 2013. *Scraping Websites to Collect Consumption and Price Data.* Available at: http://www.datakind.org/projects/food-price-scraping/ [Accessed January 2014].

De Choudury, M., Gamon, M., Counts, S., and Horvitz, E., 2013a. Predicting Depression via Social Media. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media* (Boston, MA, Jul 8-Jul 10, 2013). ICWSM 2013.

De Choudhury, M., Counts, S., and Horvitz, E., 2013b. Predicting Postpartum Changes in Emotion and Behavior via Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France, Apr 27-May 2, 2013).

Dugas, A.F., Hsieh, Y., Levin, S.R., Pines, J.M., Mareiniss, D.P., Mohareb, A., Gaydos, C.A., Perl, T.M., & Rothman, R.E. 2012.

Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics. *Clinical Infectious Diseases,* 54(4), pp.463–469.

Eagle, N., 2008. Behavioral Inference Across Cultures: Using Telephones as a Cultural Lens. *IEEE Intelligent Systems,* 23(4), pp.62–64.

Eagle, N. & Sandy Pentland, A., 2005. Reality Mining: Sensing Complex Social Systems. *Personal and Ubiquitous Computing,* 10(4), pp.255–268.

Eagle, N., de Montjoye, Y.-A. & Bettencourt, L.M.A., 2009a. Community Computing: Comparisons Between Rural and Urban Societies Using Mobile Phone Data. I*nternational Conference on Computational Science and Engineering 2009*, pp.144–150.

Eagle, N., Macy, M. & Claxton, R., 2010. Network Diversity and Economic Development. *Science,* 328(5981), pp.1029–1031.

Eagle, N., Pentland, A.S. & Lazer, D., 2009b. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America,* 106(36), pp.15274–15278.

Eysenbach, G., 2006. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. *AMIA Annual Symposium Proceedings,* pp.244-248.

Freifeld, C.C., Mandl, K.D., Reis, B.Y., & Brownstein, J.S., 2008. HealthMap: Global Infectious Disease Monitoring Through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, 15(2), pp.150–157.

Frías-Martinez, E., Williamson, G. & Frías-Martinez, V., 2011. Agent-Based Modelling of Epidemic Spreading Using Social Networks and Human Mobility Patterns. *IEEE Third International Conference on Social Computing*, pp.57–64.

Frías-Martinez, V., Soguero-Ruiz, C., Josephidou, M., & Frías-Martinez, E., 2013. Forecasting Socioeconomic Trends with Cell Phone Records. *Proceedings of the 3rd ACM Symposium on Computing for Development.*

Frías-Martinez, V., Frías-Martinez, E. & Oliver, N., 2010a. A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records. *AAAI Spring Symposium: Artificial Intelligence for Development.*

Frías-Martinez, V., Rubio, A. & Frías-Martinez, E., 2012. Measuring the Impact of Epidemic Alerts on Human Mobility. *Pervasive Urban Applications.*

Frías-Martinez, V., Virseda, J. & Frías-Martinez, E., 2010b. Socio-Economic Levels and Human Mobility. *Qual Meets Quant Workshop @ ICTD'10.*

Ford Foundation, 2012. *Ford Foundation Hosts Wired for Change*. Available at: http://www.fordfoundation.org/issues/freedom-of-expression/advancing-media-rights-and-access/news?id=677 [Accessed January 2014].

Frontline SMS, 2014. *FrontlineSMS | FrontlineCloud.* Available at http://www.frontlinesms.com [Accessed January 2014].

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., & Brilliant, L., 2008. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, 457(7232), pp.1012–1014.

Google, 2014. *Google Trends.* Available at: http://www.google.com/trends/ [Accessed January 2014].

Google, 2014. *Google Flu Trends.* Available at: http://www.google.org/flutrends/ [Accessed January 2014].

Greenberg, P., 2014. The Gates Foundation and Data. *Global Business Travel Association.* Available at: http://www.gbta.org/magazine/Pages/TheGatesFoundation+Data.aspx [Accessed January 2014].

Gutierrez, T., Krings, G. & Blondel, V.D., 2013. *Evaluating Socio-Economic State of a Country Analyzing Airtime Credit and Mobile Phone Data sets.*

Hay, S.I., Omumbo, J.A., Craig, M.H., & Snow, R.W. 2012. Earth Observation, Geographic Information Systems and Plasmodium Falciparum Malaria in Sub-Saharan Africa. *Advances in Parasitology,* 47, pp.173–215.

Humanitarian Innovation Fund, 2013. *mVAM's* Blog. Available at: http://www.humanitarianinnovation.org/blog/1442 [Accessed January 2014].

Information Week, 2013. *Gates Foundation Big Data Grants Stress Open Data*. Available at: http://www.informationweek.com/big-data/big-data-analytics/gates-foundation-big-data-grants-stress-open-data/d/d-id/1112754 [Accessed January 2014].

InSTEDD, 2006. *GeoChat | InSTEDD*. Available at: http://instedd.org/technologies/geochat/ [Accessed January 2014].

Jack, W. & Suri, T., 2011. Mobile Money: The Economics of M-PESA. *NBER Working Papers*, pp.1–30.

Kelly, M., Blanchard, S.D., Kersten, E., & Koy, K., 2011. Terrestrial Remotely Sensed Imagery in Support of Public Health: New Avenues of Research Using Object-Based Image Analysis. *Remote Sensing*, 3(11), pp.2321–2345.

Kirkpatrick, R., 2010. A Possible Role for Crowdsourcing at the United Nations? *United Nations Global Pulse*. Available at: http://www.unglobalpulse.org/blog/possible-role-crowdsourcing-united-nations [Accessed December 26, 2013].

Kramer, A.D., 2010. An Unobtrusive Behavioral Model of Gross National Happiness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp.287–290.

Krieck, M. & Dreesman, J., 2011. A New Age of Public Health: Identifying Disease Outbreaks by Analyzing Tweets. *Proc. of Health WebScience Workshop in conjunction with ACM Web Science Conference*, pp.1–5.

Krikorian, R., 2014. Introducing Twitter Data Grants. *Twitter.* Available at: https://blog.twitter.com/2014/introducing-twitter-data-grants [Accessed April 2014].

Lampos, V. & Cristianini, N., 2010. Tracking the Flu Pandemic by Monitoring the Social Web. *2010 2nd International Workshop on Cognitive Information Processing,* pp.411–416.

Leetaru, K., 2011. Culturomics 2.0: Forecasting Large-Scale Human Behavior Using Global News Media Tone in Time and Space. *First Monday,* 16(9).

Liu, B., 2012. *Sentiment Analysis and Opinion Mining,* Morgan & Claypool Publishers: San Rafael, CA.

Lopes, L.F., Zamite, J.M., Tavares, B.C., Couto, F.M., Silva, F., & Silva, M.J., 2009. Automated Social Network Epidemic Data Collector. I*NForum-Simpósio de Informática, p*p.263–272.

Lopez, G. & St. Amand, W., 2013. Discovering Global Socio Economic Trends Hidden in Big Data. *UN Global Pulse*. Available at: http://www.unglobalpulse.org/discoveringtrendsinbigdata-CHguestpost [Accessed December 2013].

Lu, X., Bengtsson, L. & Holme, P., 2012. Predictability of Population Displacement After the 2010 Haiti Earthquake. *Proceedings of the National Academy of Sciences of the United States of America,* 109(29), pp.11576–11581.

Machault, V., Vignolles, C., Borchi, F., Vounatsou, P., Pages, F., Briolant, S., Lacaux, J.-P., & Rogier, C., 2011. The Use of Remotely Sensed Environmental Data in the Study of Malaria. *Geospatial Health,* 5(2), pp.151–168.

Majewski, S., & Beger, G., 2013. *Tracking Anti-Vaccination Sentiment in Eastern European Social Networks.* UNICEF New York, Division of Communications, Social and Civic Media Section.

Mao, H., Shuai, X., Ahn, Y., & Bollen, J., 2013. Mobile Communications Reveal the Regional Economy of Côte d'Ivoire. *NetMob 2013: Data for Development Challenge.* Available at: http://www.cs.indiana.edu/~xshuai/papers/CallRank.pdf [Accessed January 2014].

Mas, I. & Radcliffe, D., 2010. Mobile Payments Go Viral: M-PESA in Kenya. *Journal of Financial Transformation,* 32, pp.169–182.

Mbiti, I. & Weil, D.N., 2011. Mobile Banking: The Impact of M-Pesa in Kenya. *NBER Working Papers.*

Meier, P., 2008. *Crisis Mapping Kenya's Election Violence.* Available at: http://irevolution.net/2008/10/23/mapping-kenyas-election-violence/ [Accessed January 2014].

Meier, P., 2012. *Crisis Mapping Syria: Automated Data Mining and Crowdsourced Human Intelligence*. Available at: http://

irevolution.net/2012/03/25/crisis-mapping-syria/ [Accessed January 2014].

Meier, P., 2013. *Using Big Data to Inform Poverty Strategies*. Available at: http://irevolution.net/2013/06/19/pulse-of-egypt-to-inform-poverty-reduction/ [Accessed January 2014].

Melville, P., Chenthamarakshan, V., Lawrence, R.D., Powell, J., & Mugisha, M., 2013. Amplifying the Voice of Youth in Africa via Text Analysis. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp.1204–1212.

Molesworth, A.M., Cuevas, L.E., Connor, S.J., Morse, A.P., & Thomson, M.C., 2003. Environmental Risk and Meningitis Epidemics in Africa. *Emerging Infectious Diseases*, 9(10), p.1287.

Morrow, N., Mock, N., Papendieck, A., & Kocmich, N., 2011. *Independent Evaluation of the Ushahidi Haiti Project, Development Information Systems International/Ushahidi Haiti Project.*

Munro, R., 2010. Crowdsourced Translation for Emergency Response in Haiti: the Global Collaboration of Local Knowledge. *AMTA Workshop on Collaborative Crowdsourcing for Translation*, pp.1–4.

mWater, 2013. *USAID Invests in mWater for Social Water Monitoring*. Available at: http://mwater.co/news/div/ [Accessed January 2014].

Nielsen, R.C., 2013. Exhibiting the Global Post-2015 Twitter Conversation. *UN Global Pulse.* Available at: http://www.unglobalpulse.org/Post2015Twitter [Accessed December 2013].

Noor, A.M., Alegana, V.A., Gething, P.W., Tatem, A.J., & Snow, R.W. 2008. Using Remotely Sensed Night-Time Light as a Proxy for Poverty in Africa. *Population Health Metrics,* 6(5).

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A., 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Tepper School of Business.*

Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., & Goss, C. H., 2011. Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends. *PloS One*, 6(4), e18687.

Pak, A. & Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh Conference on International Language Resources and Evaluation LREC'10.*

Paltoglou, G. & Thelwall, M., 2012. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4), p.66.

Pang, B. & Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), pp.1–135.

Paul, M.J. & Dredze, M., 2011a. *A Model for Mining Public Health Topics from Twitter.* Johns Hopkins University. Report.

Paul, M.J. & Dredze, M., 2011b. You Are What You Tweet: Analyzing Twitter for Public Health. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media.*

Pentland, A., Lazer, D., Brewer, D., & Heibeck, T., 2009. Improving Public Health and Medicine by Use of Reality Mining. *Studies in Health Technology Informatics*, 149, pp.93–102.

Pentland, A., Reid, T.G. & Heibeck, T., 2013. *Big Data and Health: Revolutionizing Medicine and Public Health, WISH Big Data and Health* report.

Pevaiz, F., Pervaiz, M., Rehman, N. A., & Saif, U., 2013. FluBreaks: Early Epidemic Detection From Google Flu Trends. *Journal of Medical Internet Research*, 14(5), e125.

Positive Innovation for the New Generation (PING), 2012. *Disease Surveillance & Mapping.* Available at: http://www.pingsite.org/tech-projects/disease-surveillance-project/ [Accessed January 2014].

Power, R., Chen, J., Kuppusamy, T. K., & Subramanian, L., 2010. Document Classification for Focused Topics. *AAAI Spring*

Symposium: Artificial Intelligence for Development.

Preis, T., Moat, H. S., Stanley, H. E., & Bishop, S. R., 2012. Quantifying the Advantage of Looking Forward. *Scientific Reports*, 2. doi:10.1038/srep00350.

Quinn, J.A., Okori, W. & Gidudu, A., 2010. Increased-Specificity Famine Prediction Using Satellite Observation Data. *Proceedings of the First ACM Symposium on Computing for Development.*

Radojevic, M.B., 2012. *Social Media and Political Risk Analysis,* UNDP in Central Europe and Asia. Available at: http://europeandcis.undp.org/blog/2012/11/16/social-media-and-political-risk-analysis/ [Accessed January 2014].

Rahman, A., Kogan, F. & Roytman, L., 2006. Short Report: Analysis of Malaria Cases in Bangladesh with Remote Sensing Data. *American Journal of Tropical Medicine and Hygiene*, 74(1), pp.17–19.

Randrianasolo, L., Raoelina, Y., Ratsitorahina, M., Ravolomanana, L., Andriamandimby, S., Heraud, J.-M., et al., 2010. Sentinel Surveillance System for Early Outbreak Detection in Madagascar. *BMC Public Health,* 10:31.

RapidSMS, 2013. *RapidSMS Home Page*. Available at: http://www.rapidsms.com. [Accessed January 2014].

Rigobon, R., 2011. *Bread Price Index: a Pilot Case*. Technical Note.

Royal Tropical Institute, 2013. *mHealth Projects: Examples from Low- and Middle-Income Countries.* Available at: http://www.mhealthinfo.org/projects_table [Accessed December 2013].

Sadilek, A., Kautz, H.A. & Silenzio, V., 2012. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence.*

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E.M., Cattuto, C., Khandelwal, S., Mabry, P.L., & Vespignani, A., 2012. Digital Epidemiology. *PLoS Computational Biology*, 8(7), e1002616.

SAS & UN Global Pulse, 2011. *Using Social Media and Online Conversations to Add Depth to Unemployment Statistics: Methodological White Paper,* Available at: http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics.

Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J., 2010. The Utility of "Google Trends" for Epidemiological Research: Lyme Disease as an Example. *Geospatial Health,* 4(2), pp.135–137.

Shrivastava, A. & Pant, B., 2012. Opinion Extraction and Classification of Real Time Facebook Status. *Global Journal of Computer Science and Technology,* 12(8).

Skoll World Forum, 2013. *About Skoll World Forum.* Available at: http://skollworldforum.org/about/ [Accessed January 2014].

Soto, V., Frías-Martinez, V., Virseda, J., & Frías-Martinez, E.,2011. Prediction of Socioeconomic Levels Using Cell Phone Records. In *User Modeling, Adaptation and Personalization*, pp.377–388. Springer.

Stieglitz, S. & Dang-Xuan, L., 2012. Political Communication and Influence through Microblogging — An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior. *Proceedings of the 2012 45th Hawaii International Conference on System Sciences,* pp.3500–3509.

Tatem, A.J. & Hay, S.I., 2004. Measuring Urbanization Pattern and Extent for Malaria Research: a Review of Remote Sensing Approaches. *Journal of Urban Health,* 81(3), pp.363–376.

Telefónica, 2014. *Telefónica.* Available at: http://telefonica.com/es/home/jsp/home.jsp [Accessed January 2014].

Thomson, M.C. & Connor, S.J., 2001. The Development of Malaria Early Warning Systems for Africa. *Trends in Parasitology,* 17(9), pp.438–445.

Tromp, E., 2012. *Multilingual Sentiment Analysis on Social Media.* Eindhoven University of Technology.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M., 2010. Predicting Elections with Twitter: What 140 Characters Reveal

About Political Sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media,* 178–185.

UN Global Pulse, 2013a. Advocacy Monitoring through Social Data: Womens and Childrens Health. *UN Global Pulse*. Available at: http://www.unglobalpulse.org/node/14661 [Accessed December 2013].

UN Global Pulse, 2013b. Big Data for Development: Challenges and Opportunities.

UN Global Pulse, 2013c. Daily Tracking of Commodity Prices: The e-Bread Index. *UN Global Pulse.* Available at: http://www.unglobalpulse.org/projects/comparing-global-prices-local-products-real-time-e-pricing-bread [Accessed December 2013].

UN Global Pulse, 2013d. Global Snapshot of Wellbeing - Mobile Survey. *UN Global Pulse.* Available at: http://www.unglobalpulse.org/projects/global-snapshot-wellbeing-mobile-survey [Accessed December 26, 2013].

UN Global Pulse, 2013e. *Mobile Phone Network Data for Development,* Available at: http://www.slideshare.net/unglobalpulse/mobile-data-for-development-primer-october-2013.

UN Global Pulse, 2013f. Twitter and Perceptions of Crisis-Related Stress. *UN Global Pulse*. Available at: http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress [Accessed December 2013].

UN Global Pulse, 2013g. Unemployment Through the Lens of Social Media. *UN Global Pulse.* Available at: http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics [Accessed December 2013].

UN Global Pulse, 2013h. Question Box Analytics. *UN Global Pulse.* Available at: http://www.unglobalpulse.org/projects/question-box-analytics [Accessed April 2014].

UN Global Pulse, 2014. Analyzing Online Content for Insight on Women and Employment in Indonesia. *UN Global Pulse.* Available at: http://www.unglobalpulse.org/indonesia-women-employment [Accessed April 2014].

UNICEF Uganda, 2014. *UReport - Voice Matters.* Available at: http://ureport.ug/ [Accessed January 2014].

UNICEF, WFP, & UN Global Pulse, 2012. *Food and Nutrition Security Monitoring and Analysis Systems.*

Union of Concerned Scientists (UCS), 2013. *UCS Satellite Database*. Available at: http://www.ucsusa.org/nuclear_weapons_and_global_security/space_weapons/technical_issues/ucs-satellite-database.html [Accessed January 2014].

Valdivia, A., Lopez-Alcalde, J., Vicente, M., Pichiule, M., Ruiz, M., & Ordobas, M., 2010. Monitoring Influenza Activity in Europe with Google Flu Trends: Comparison with the Findings of Sentinel Physician Networks-Results for 2009-10. *Eurosurveillance,* 15(29), 2–7.

Vital Wave Consulting, 2009. *mHealth for Development: The Opportunity for Mobile Technology for Healthcare in the Developing World,* UN Foundation/Vodafone Foundation.

Watts, G., 2009. Google Watches Over Flu. *British Medical Journal,* 338, pp.74–75.

Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., & Buckee, C.O., 2012. Quantifying the Impact of Human Mobility on Malaria. *Science*, 338(6104), 267–270. doi:10.1126/science.1223467.

World Bank & DataKind, 2013. Final Report. *DC Big Data Exploration* (Workshop).

World Food Programme (WFP), 2013. *"Press 1 If You Did Not Eat Yesterday...".* Available at: http://www.wfp.org/stories/press-1-if-you-did-not-eat-yesterday-congo [Accessed January 2014].

World Health Organization (WHO), 2011. *mHealth: New Horizons for Health Through Mobile Technologies: Second Global Survey on eHealth*, World Health Organization.

Wu, L., 2009. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. *ICIS 2009 Proceedings.*