# How Big Data Reshapes Knowledge for International Development – A Governmentality Perspective

Mikkel Flyverbom, Anders Koed Madsen and Andreas Rasche

Copenhagen Business School

Porcelænshaven 18A, 2000 Frederiksberg, Denmark

Working Draft – EGOS 2016, comments welcome: ara.ikl@cbs.dk

**Abstract:**
The aim of this paper is conceptualize and illustrate how large-scale data and algorithms condition and reshape knowledge production when addressing international development challenges. Based on a review of relevant literature on the uses of big data in the context of development, we unpack how digital traces from cell phone data, social media data or data from internet searches are used as sources of knowledge in this area. We draw on insights from governmentality studies and argue that big data's impact on how relevant development problems are governed revolves around (1) new techniques of visualizing development issues, (2) a reliance on algorithmic operations that synthesize large-scale data, (3) and novel ways of rationalizing the knowledge claims that underlie development efforts. Our discussion shows that the reliance on big data challenges some aspects of traditional ways to collect and analyze data for development (e.g. via household surveys and deductive approaches), and we articulate intersections between different kinds of knowledge production, different ways of collecting and controlling data, and different epistemic foundations for addressing and governing development problems.

## INTRODUCTION

The masses of digital data produced by internet traffic (e.g. Google searches, tweets, Facebook posts) and various forms of tracking and navigation (e.g. GPS devices, mobile logs) offer new insights into human practices and hidden societal trends. The term "big data" has been used to describe such large datasets requiring new forms of data storage, analysis and visualization technologies (Chen et al., 2012). The excitement surrounding big data is both about the existence of larger volumes of data and the ability to aggregate, search, and cross-reference these volumes (Boyd and Crawford, 2012). Early uses and discussions of big data focused on marketing and other corporate uses, such as gathering real-time data about customer behavior to develop targeted advertising and improved decision-making (McAfee and Brynjolfsson, 2012). In recent years we have seen subsets of research that direct analyses towards understanding other uses of big data. One example is the emerging literature on Big Data for development (BD4D), which explores the potential of big data to produce actionable information that can be used to improve development, e.g. by identifying needs, providing services, and predicting crises (Ginsberg et al., 2009).

The aim of this paper is to articulate how big data analyses condition and contribute to international development. We will survey the field of BD4D and use the analysis of two specific big data projects to provide insights into the more general mechanisms by which big data renders certain areas of international development governable. The theoretical framework used to address this question is based on Michel Foucault's (1978, 1991a) notion of governmentality and, in particular, its discussion through the work of Mitchell Dean (1996, 2009). Governmentality aims "to uncover and examine the often invisible rationality which is behind an assemblage of actions and mechanisms that are in place to govern certain actions" (Gouldson and Bebbington, 2007: 12). Although scholarly work has used the lens of governmentality to explore the rationalities, practices and power structures underlying

international development (Methmann, 2011; Murray, 2007), we still know little about how changes in sourcing, processing and communicating data affect the way development problems are addressed. This is a surprising omission since data and the analytical techniques attached to it shape how governance problems are "re-presented in the place where decisions are to be made about them" (Miller and Rose, 1990: 7).

The governmentality lens fits well with our research aim because it offers a theoretical framework committed to grasping the specificities and material conditions that need to be in place in order for big data to shape the production of knowledge for international development. We structure the discussion around Dean's (2009) dimensions of an analytics of government: the fields of visibility surrounding regimes of practices; the instruments and techniques (*techne*) that enable and constrain these regimes; and the forms of knowledge (*episteme*) attached to certain regimes. These dimensions provide a framework to examine the specific conditions under which BD4D emerges as a way of processing data to address development challenges. The identification of these conditions is central to our goal, namely to articulate how reliance on digital data and algorithms both extends and challenges international development efforts.

Using Dean's dimensions in our analysis, we argue that the uptake of BD4D is conditioned upon: (1) the utilization of new ways of visualizing development problems and hence exposing unacknowledged dimensions of these problems; (2) an acceptance by development organizations of the impossibility of producing and controlling data "in-house", instead relying on data provided by a distributed set of private companies; and (3) the acceptance of new epistemic foundations for governing development problems. Taken together, these issues problematize selected aspects of existing data-processing practices used in the field of development and question established assumptions about the knowledge-power nexus that guides international development. Our analysis does *not* argue that big data

replaces traditional ways of handling data in international development, nor do we argue that BD4D is without problems. BD4D opens new perspectives on the handling of development data and development issues.

The contribution of this paper is twofold. First, we extend the emerging scholarly discourse on the societal relevance of big data (see e.g. Hilbert, 2013) by theorizing BD4D as a particular, albeit subtle form of governmentality, shaping how problems and opportunities are framed and acted upon. Second, we contribute to discussions on how to govern social, environmental and economic affairs in the field of development. While it is widely recognized that the international development agenda is shifting (Ilcan and Phillips, 2010), the effects of working with large-scale digital data remain under-explored.

## BIG DATA FOR DEVELOPMENT

### Characterizing big data

The possibility of using big data as indices of social dynamics is rooted in a widespread "datafication" (Mayer-Schönberger and Cukier, 2013: 73) whereby multiple elements of social life are quantified and stored in the shape of digital data. The resulting indices have their own district characteristics when compared with, for instance, deductive statistics that usually guide projects in international development. A popular account has suggested that it is the growing *volume*, *velocity* and *variety* of data that makes big data different from previous modes of data analysis (Laney, 2001). Others have downplayed this emphasis on volume and pointed to the possibility of integrating and aggregating fast-moving and disparate forms of data through new types of databases and calculative software (Boyd and Crawford, 2012). We consider the importance of big data for knowledge and governance to

be less about the nature of data and more about how the analytical processes and algorithmic operations involved guide our attention in particular ways (Amoore and Piotukh, 2015).

Few discussions start from a clear-cut definition of what big data is and is not. However, in order to conceptualize its consequences for international development we need a minimal definition of the phenomenon. We focus on the intersection of widespread datafication (Mayer-Schönberger and Cukier, 2013) and algorithmic developments allowing for the mining, correlation and visualization of digital traces from disparate sources (Gitelman, 2013). Algorithms are, in this context, understood as generalized procedures for turning dis-organized data-inputs into manageable outputs through series of logical rules that provide instructions on how to handle data with specific attributes. Along these lines, we define big data as *algorithm-based analyses of large-scale, disparate digital data that is collected for purposes of prediction, measurement and governance*.

**Approaching Big Data for Development (BD4D)**

International development encompasses a wide range of approaches, activities and institutions seeking to improve the quality of life of people around the world (Sumner and Tribe, 2008). This includes initiatives targeting both "economic and social development and encompasses many issues such as humanitarian and foreign aid, poverty alleviation, the rule of law and governance, food and water security, capacity building, healthcare and education, women and children's rights, disaster preparedness, infrastructure, and sustainability" (Greiman, 2011: 8). In future, experiments with big data projects are likely to shape these areas, and this is why a deeper understanding of the analytical techniques and calculative rationalities involved is valuable.

We investigate how particular forms of knowledge production and indices guide development efforts. This link between indices and governance is well established in the sociological literature on numbers. For instance, Merry (2011) argues that: "Indicators are a technology of not only knowledge production but also governance. They are widely used for decisions, such as where to send foreign aid, where to focus on human rights violators, and which countries to offer the best conditions for business development." Along similar lines of thinking, we conceptualize big data as a form of knowledge production with governance effects. Development issues have historically been approached and accounted for through numbers and statistics. Such indicators include, most famously, the United Nations Development Program's yearly *Human Development Report*, which revolves around indices such as the Human Development Index, the Gender Development Index and the Human Poverty Index (Sumner and Tribe, 2008). Such indicators rely on national statistics such as per capita income, household surveys and similar numbers-based types of data.

Besides such indices, international organizations often turn to survey-based data in order to create, monitor and evaluate development policies (Ginsberg et al., 2009; United Nations, 2013). The UN and the World Bank use household surveys for collecting information on populations in developing and emerging economies (World Bank, 2004; United Nations, 2005). Samman (2013) argues that such surveys are still the main workhorse of data collection for international development; however, she also recognizes that there are limits to survey-based data collection. Surveys are often restricted to the heads of households, making it difficult to gather data on some topics (e.g. in-home violence). Data collection through household surveys is also costly, and results are usually only available after some delay (Deaton, 2000). Although some of these aspects are changing due to policy priorities (e.g., the Sustainable Development Goals call for gathering disaggregated data) and the use of

new devices (e.g., the use tablet computers), data gathered through household surveys can still be complemented in a number of ways.

BD4D has been introduced to *supplement* survey-based data in selected contexts where new types of data have become accessible. The production of large amounts of data is no longer restricted to the industrialized world. Driven by regulatory reforms of telecommunication sectors, high investment activity and decreasing costs of adoption (Howard and Mazaheri, 2009), some developing countries have witnessed rising internet and mobile phone usage. Countries in sub-Saharan Africa have a mobile phone penetration rate of more than 60% (GSMA, 2012) as mobile technologies are often used as a substitute for weak cable infrastructures. Mobile phones are routinely used to transfer money, search for work, and to transmit medical information. Data generated through such usage can produce information that is relevant when addressing development challenges. Many developing and emerging countries have also witnessed rapid growth of social media (Kohut et al., 2011). Although the uptake of relevant technologies remains uneven among countries (Chinn and Fairlie, 2010) and there still is a "digital divide" in some respects (International Telecommunication Union, 2012), there is a growing awareness of the potential in using big data to address development problems.

**A review of existing literature on BD4D**

While not exhaustive, this review of the existing literature on BD4D shows that the field is dominated by two tendencies: one is that the primary data sources are cell phones, internet searches and social media sites; the other is that the most prevalent uses of these sources are real-time monitoring of physical movements and faster predictions of future events in various contexts.

***Cell phone data as an indicator of movement and migration.*** Studies of BD4D projects involving cell phone data have shown such data to be particularly useful in the management of challenges associated with human movement and migration. Bayir et al. (2009) and Deville et al. (2014) have described how cell phone data is used to understand and react to mobility-patterns internally in countries or regions. Also, Blumenstock (2012) has shown how call records have been used to spot patterns of migration in Rwanda that were unobservable using standard survey techniques. Other scholars have studied the use of cell phone data to tackle problems related to disease outbreaks. Wesolowski et al. (2012) address such use in relation to the spread of malaria in Kenya, while Frías-Martínez et al. (2011) have conducted similar work on the 2009 H1N1 outbreak in Mexico. Others have focused on the uses of cell phone data in the face of natural disasters. For instance, Morales et al. (2015) have studied the use of call records data to characterize human mobility during floods, while Moumni et al. (2013) have taken earthquakes in Mexico as their empirical case. Finally, we have seen scholars focusing on how the act of mobile phone recharging can be used as a real-time indicator of the level of socio-economic development in specific regions (Letouzé, 2012; Frías-Martínez and Virseda, 2013).

***Internet searches as indicators of physical and economic wellbeing.*** Internet searches are used to develop the kinds of predictive analytics that are part of many big data projects. In the area of disease prevention, Althouse et al. (2011) have described how Google searches have been used to predict dengue outbreaks in Singapore and Thailand before traditional data sources could obtain relevant signals. However, diseases are not the only threat to wellbeing that internet searches have been used to predict. Searches related to issues of unemployment correlate with the unemployment levels disclosed by the U.S. government and are faster indices of economic wellbeing than statistical sources (Hubbard, 2011; Ettredge et al., 2005). We have also seen studies of initiatives that use internet searches as indices in relation to

dynamics of human reproduction. Reis and Brownstein (2010) have shown that volumes of internet searches for abortion are directly proportional to local restrictions on abortion. Similarly, Billari et al. (2013) have shown that Google searches for fertility-related queries like "pregnancy" or "birth" can be used to predict fertility intentions and fertility rates several months ahead.

***Social media traces as multi-purpose indicators***. Social media sites are a diverse data source in terms of the specific platforms that deliver data, the kind of data they deliver and the uses to which they are put. Such platforms include Twitter, Facebook, Flickr and Foursquare, which produce unstructured data in the form of text and dialogues and more structured metadata such as geo-tags and time stamps. Traces of social media interactions have been used as indices in relation to a broad range of development problems. A number of studies have investigated how social media data can be used to track movements: De Choudhury et al. (2010) have reconstructed travel-patterns on the basis of geo-tagged pictures on Twitter; Ferrari and Mamei (2011) have discovered people's whereabouts with Google Latitude; and Ferrari et al. (2011) have identified mobility patterns in urban areas through location data in tweets. Studies on the use of traces from social media as predictive indices are also numerous. Chunara et al. (2012) have described how Twitter posts were used to spot relevant signals of disease in the 2010 Haitian cholera outbreak two weeks before the official data. Ritterman et al. (2009) have undertaken similar studies in relation to predictions of Swine Flu Pandemics. Outside of the field of epidemiology, we have seen studies of Twitter-based predictions of, amongst other things, criminal incidents **(**Wang et al., 2012) and touristic seasonal patterns (Mocanu et al., 2013).

To guide and organize our analyses of how such uses of big data condition knowing and governing, we rely on an analytical framework based on the governmentality literature.

## GOVERNMENTALITY: EXPLORING CONDITIONS FOR KNOWING AND GOVERNING

The concept of governmentality builds on and extends Foucault's work, in particular how power and control take the shape of "actions on other's actions" that structure and normalize particular ways of acting and thinking (Foucault, 1982). What Foucault (1991a) terms government reaches beyond the established meaning of the word and involves all sorts of attempts to steer human conduct. The governmentality literature offers an analytical vocabulary that is useful if we want to understand and conceptualize the emergence of particular ways of governing conduct, in our case how big data analyses condition and guide development efforts. The main strength of a governmentality perspective is that it views all attempts to govern conduct as a complex of operations, calculations and reasonings rather than direct interventions by the state or other actors. This also implies that governmentality approaches view power not as something that is allocated by means of structural properties (e.g. state policies) but as much more fuzzy and entwined attempts to exercise power "at a distance" (Miller and Rose, 1990). Rather than focus on the resources or positions of actors seeking to shape conduct, a governmentality approach investigates how practices and their underlying techniques and rationalities condition particular forms of governance.

One extension of Foucault's thoughts on governmentality is Dean's "analytics of government" framework, which sets out "to show the conditions under which regimes of practices come into being, are maintained and transformed" (Dean 2009: 30). Regimes of practices can be understood as organized patterns by which we do things, such as collecting data or assessing needs. Regimes of practices are often taken for granted, play important roles in institutional and organizational settings, and are referred to as "systems" (e.g. data

processing systems) since they cannot be ascribed to single individuals or groups. Our analysis is concerned with big data as a regime of practices for collecting and analyzing data relevant to international development.

Dean (1996, 2009) has introduced an analytics of government as a framework with which to critically analyze the conditions under which regimes of practices operate. The first dimension, *fields of visibility*, centers on what is rendered visible and invisible, i.e. made knowable and thus governable, in governance efforts. The second dimension, *technes of government*, focuses on the manifold mechanisms and techniques used to exercise governance. The third dimension, *epistemes of government*, considers the rationalities and worldviews that underpin a particular regime of practices. These dimensions are co-present within regimes of practices and presuppose one another. We use these dimensions to analyze how big data analyses condition governance efforts in the area of international development. How governance efforts position human subjects and shape identities is an important focus of governmentality approaches, but the data we rely on does not allow us to address this question. However, as we state in the conclusion, these are important issues to examine in future research.

## ANALYZING BIG DATA FOR DEVELOPMENT AS GOVERNMENTALITY

Our analysis examines two big data projects that cover two of the main data sources mentioned above as well as the focus on monitoring and predicting which the literature review showed to be central. The first case concerns the *use of mobile data to track human behavior* and focuses on regional movements within Kenya. On the basis of data from the market-leading mobile carrier, a group of researchers have illustrated the potentials of big data in the optimization of disease prevention (Weslowski et al., 2012, 2013). They show

how mobile data can be used to visualize the dynamics of human carriers of malaria and distinguish between regions that are respectively sources and sinks of this disease. Furthermore, they illustrate how mobile data can enrich data from human population censuses by adding short-term travel patterns to the yearly migration patterns depicted through censuses. The second case illustrates the use of *tweets to track social phenomena in real time*, such as when the UN Global Pulse relies on tweets to track food prices and food crises (Letouzé, 2012; UN Global Pulse and Crimson Hexagon, 2011). Through a semantic analysis of Twitter content, the Lab has shown that sentiments in tweets can be an early warning of economic crises in situations where official statistics are produced with a time lag.

**The Field of Visibility Characterizing BD4D**

There are many ways to make objects of government visible: flow charts, data tables, maps and organograms are just some examples (Dean, 2009: 41). Ways of visualizing simultaneously define objects of government, since they highlight some characteristics while hiding others. This is particularly true given that a great deal of knowledge in and between organizations assumes the form of visual representations (e.g. through PowerPoint). Visualizations may even obscure certain objects completely (e.g. when a flow chart omits a process). Dean stresses that a field of visibility involved in a regime of practice conditions who and what is to be governed, which governance problems are to be addressed and with what objectives in mind.

When looking at the visual outputs of the two cases it is evident that the creation of new visibilities is a core aspect of BD4D: they produce visualizations to improve the gaze of BD4D practitioners. Figure 1 visualizes regional movements in Kenya with color codes indicating the intensity of movements in specific regions. Blue-colored regions have many

people leaving, whereas green-colored areas denote regions where people rarely leave. Since this visualization is based on mobile data with 20-minute intervals it provides information in a much more dynamic manner compared to similar visualizations relying on census data. Figure 2 shows how Twitter-based visualizations can provide new types of visibilities. UN Global Pulse has used the visualization technique of word-clouds to build a Twitter-based crisis monitor that indicates how people are impacted by food prices in Indonesia. These word-clouds get their specific visual shape by balancing human and algorithmic inputs into the data processing. Initially the UN used experts to "train" the algorithm to recognize specific themes and emotions. However, in order to handle the speed of data these word-clouds are then shaped by an automated algorithmic recognition of patterns in the semantic content of tweets (UN Global Pulse and Crimson Hexagon, 2011). Such balance is the foundation for transforming real-time data feeds from Twitter into word-clouds, semantic clusters and color-coded topics that make weak signals of economic crises visible in real time (e.g. when the non-affordability of food impacts on living conditions).

==========

Insert Figure 1 and 2

==========

The cases show that BD4D as a regime of practice produces certain forms of visibility. These range from traditional ways of depicting development data (e.g. colored geo-maps) to new visualization techniques (e.g. word-clouds). Data visualization techniques make development problems governable in novel ways. Our argument is not simply that big data analyses offer better depictions but also that they produce new visibilities and targets of governance. *What* problems are being made visible through BD4D depends to some degree on *how* the underlying data is visualized and *who* is captured in the visualizations. The "how"

of visualization is important because choices concerning the logics of categorization beneath the visual representations are controversial. The priority for the UN to give topical experts a central role in "training" the algorithm is, for instance, a priority that serves to maintain a certain order of knowledge within the organization. Human experts' categories and rationalities are still the foundation from which algorithms are produced and evaluated. Such choices concerning the balance between algorithmic intelligence and human expertise are central when producing actionable visualizations for developmental purposes. Similarly, the "who" of visualization has implications, as the probability of being surveyed by traditional methods is rather low compared to a person who tweets or sends location data.

The visualizations associated with BD4D offer new perspectives on information contained in data sets that would otherwise be hard to grasp. The crisis monitor depicted in Figure 2 not only organizes unstructured Twitter data into identifiable topics related to food prices but also allows users to explore the stories and tweets behind certain topics in real-time. In both cases the focus on speed in the production of visualizations distinguishes BD4D from classic statistics like household surveys. Also, visualizations based on digital traces enable gradual shifts between levels of analysis. With digital data it has become increasingly possible to zoom in and out between an aggregated body of data and the individual context around a specific source of data (Latour et al. 2012). For instance, it is easy to follow a link between two words in an aggregated word cloud back to the set of tweets that serves as the basis for this link. From there it is possible to explore other metadata around these tweets.

**The Technical Aspects of Governing Through BD4D**

What Dean (2009: 42) calls the *techne* of government relates to the manifold mechanisms and techniques by which authority is exercised. The governmentality literature uses the term to refer to mechanisms like rankings that normalize behavior, compliance mechanisms like codes of conduct, and techniques of education. Such "humble and mundane mechanisms, which appear to make it possible to govern" (Miller and Rose, 1990: 8) are worth exploring because they turn reality into something that is governable and can be acted upon.

We argue that it is impossible to understand the relation between big data and governance practices without inquiring into the technologies underpinning the visibilities discussed above. We suggest that the practice of governing international development through big data is inevitably shaped by the *affordances* of the computational technologies used. The affordance of a technology denotes the way in which it supports or limits certain uses in a given context (Gaver, 1991). To understand the ways in which BD4D shapes international development we need to consider the means by which digital traces are sourced, aggregated and visualized. When looking at our cases, two types of technologies are decisive: (1) the technologies they use to harness and quantify data, or what we call *sourcing software*; and (2) the technologies providing the formal instructions needed for a computer to transform inputs from the sourcing software into visual outputs, i.e. *synthesizing algorithms*.

***Sourcing software.*** Both cases illustrate how the affordances of sourcing software come to condition the types of knowledge BD4D projects can produce. For instance, the choice to track human movements through mobile data means that the visualizations produced are shaped by the infrastructure and market penetration of the mobile carrier that delivers the data. The distance between the routing towers of a mobile carrier impacts the quality of a given visualization since shifts in a person's tower connection are used as an indication of his or her movement. Similarly, the market penetration of the mobile carrier is

important because the customer base will determine the scope of the population made visible. In the case of Kenya there is a single carrier with a market share of 92%; however, such a near monopoly does not exist in all countries. BD4D projects based on mobile data may therefore either have a lower generalizability or require a combination of data from different carriers with different ways of collecting data.

The reliance on sourcing software owned by private companies like Google imposes important constraints on the way BD4D-related phenomena can be knowable and governable. This also applies to projects that source data from Twitter's streaming Application Programming Interface (API). This API is the "pipeline" through which Twitter offers external partners an opportunity to interact with their database. Such APIs set up important constraints in terms of the access they provide (Boyd and Crawford, 2012). Only a few people have access to the full "firehose" of tweets and most will have to work with a smaller selection. It is not entirely transparent how that selection is sampled and whether it is censored by Twitter before being made available through the API. In other words, central methodological choices regarding sampling logics and data provenance are hard to control for organizations working with Twitter's API.

This issue of control is part of the reason why the UN's Global Pulse has partnered with an approved reseller of Twitter's data called Crimson Hexagon. By sourcing its data through this partner the UN gains more control over the software used to harness data. However, other parts of the sourcing software still set important conditions, in particular because data formats and meta-data structures are completely controlled by Twitter. Twitter's interface produces data that cannot be longer than 140 characters, can be left anonymously, and contains meta-data such as hashtags (#) and re-tweets (@). When the UN Global Pulse repurposes Twitter's data it chooses to work with short, condensed messages from a specific communicative culture that shapes the way a situation can be datafied (Rogers, 2013). The

design of data interfaces and the formats of data have a significant influence on the kind of knowledge produced by BD4D.

***Synthesizing algorithms***. Synthesizing algorithms are central to the further processing of the data generated by the sourcing software. For instance, when Twitter's API returns data in spreadsheets there is a need for algorithms that give instructions as to how the data is processed, organized and visualized (Neuhaus and Webmoor, 2012). Both cases depend on synthesizing algorithms to turn quantified data from the sourcing software into visual representations. The "outgoing ranks" that are used to color the map in Figure 1, for instance, are based on a series of data points from the mobile towers that users interact with when they use their phones. These data points show the locations of each mobile user at specific points in time. This data is then run through an algorithmic operation that aggregates the towers to country-level groups and assigns the individual mobile user a score when he/she moves from one group to another. The conclusions about Nairobi as a sink for malaria and the designations of specific settlements as risk regions were derived via a clustering algorithm working on network data. In the case of the UN Global Pulse's crisis monitor in Figure 2, algorithms are used to compute semantic distances between words in tweets on the basis of expert inputs, theories of natural language processing and network characteristics, and to turn these into semantic networks.

BD4D involves some degree of displacement of power from topical experts (who draw on past experience and established classifications) to computer scientists (who rely on machine intelligence and automatization). BD4D is partly inspired by theorists like Anderson (2008), who proposes that automated algorithms can be used to circumvent biases in the classificatory schemes of humans (e.g. hypothesis-testing). The argument is that whereas such classifications are polluted by routines, special interests and power struggles, automated algorithmic pattern detection is not. BD4D applications may also be grounded in expert

classification (King and Powell, 2008), but the focus on real time data clearly increases the importance of algorithms in international development. The formal instructions for data processing that are encoded into algorithms transfer assumptions from theories about networks and natural language to the practice of governance. Accordingly, algorithmic code will be a source of power that resembles the role classifications have played in previous regimes of governance (Bowker and Star 2000; Gillespie, 2014).

**Forms of Knowledge in BD4D**

*Episteme* relates to a concern with the worldviews that underpin regimes of practices (Dean, 2009: 42). The *episteme* of government looks at how specific forms of truth and knowledge arise from and guide regimes of practices, and how these render reality governable. As Foucault (1991b: 79) emphasized, "'practices' don't exist without a certain regime of rationality", and analyzing such rationalities implies that we study how regimes of practices justify knowledge and try to establish truth claims. The analysis of *epistemes* seeks to articulate the worldviews and assumptions about knowledge at work in governance, such as the truth claims that BD4D gives rise to. Being built from a different basis than, for instance, household surveys, big data changes how knowledge is rationalized and hence creates a different ground upon which to evaluate "truth". We argue that this rationalization differs from more established forms of knowledge production in three ways.

***Knowledge for development based on masses of behavioral data***. One of the rationales underpinning BD4D projects is that their knowledge claims are justified through reference to the *amount* of data collected. Household surveys rely on rather small amounts of sampled data because collection is time and resource-intensive (Deaton, 2000), and they define up front what kind of data is needed and how it is to be used. BD4D applications rest on an

inductive and (semi-)automated analysis of large amounts of unstructured data. Looking at the use of mobile data to track movements in Kenya we can see that this focus on quantity and induction is one of the central arguments for the legitimacy of the project. Weslowski et al. (2012) argue that the use of call data is a way to solve epistemic problems stemming from two shortcomings in other available data sources on movement: (1) data on movement is often collected for specific purposes (e.g. building a new road) which reduces its generalizability and makes it impossible to have continuous data over time; (2) data on human movement is frequently based on the survey answers of a few individuals and respondents often misinterpret what they are asked about.

The use of mobile data as indicators of movements in Kenya shows how the increased scale of data has consequences for how knowledge claims are legitimized. When using small amounts of data, knowledge claims are legitimized by pointing to the appropriateness of the underlying sample (i.e., we can learn something generalizable by looking at a randomized sample). The legitimacy attached to BD4D applications rests on using much larger samples. The fact that an analysis rests on millions of behavioral data-points, instead of a few hundred answers to a pre-defined survey, makes the resulting knowledge *appear* legitimate and capable of bypassing traditional problems such as researcher bias in the way surveys are formulated and response-bias in the way they are answered.

***Knowledge for development based on messy but real-time data.*** When creating knowledge for development by increasing the scale of samples it is necessary to move beyond clean, carefully designed data and to accept some messiness (Mayer-Schönberger and Cukier, 2013). While surveys operate on the assumption that data needs to be as correct as possible (since the sample size is limited), it is often argued that BD4D can afford some inaccuracies (e.g. the inclusion of *some* irrelevant tweets) in exchange for the benefit of analyzing a much larger dataset. This has consequences for the authority and use of

knowledge. BD4D is often not as accurate as survey methods and hence can hardly be used as a standalone strategy when approaching development problems. Its legitimacy cannot be established on the basis of claims about representativeness, as is the case with survey samples. BD4D needs to be evaluated through different epistemic standards if it is to appear as a legitimate source of knowledge.

Much of the authority given to knowledge resulting from big data analyses rests on the temporality of data. BD4D is considered a good way to identify early "digital smoke signals" (Lohr, 2013) – i.e. pulses, anomalies and trends that survey-based methods rarely capture or only capture too late. These weak signals can then be used to investigate a problem further. In relation to the use of mobile data in Kenya, for instance, a central part of the justification for working with call data was that it comes in 20-minute intervals. Therefore it provides a temporal granularity that far exceeds data-collection methods such as the annual national census. Accordingly, an important aspect of the legitimization of call data is that it allows for inquiring into the relation between short-term movements and long-term movements in a way other data sources do not.

The immediate availability of results through BD4D applications strengthens the authority of knowledge claims and offers the possibility of adjusting policies more quickly, allowing those in charge of development work to see what is (not) working (Piotrowski, 2013). This is also part of the argument justifying the way the UN Global Pulse uses Twitter data to detect early signals of crises. Negative tweets about prices of rice are available far ahead of official price statistics. Similarly, another analysis of Twitter data in Indonesia revealed that people were saying that vaccines were not halal because they contain pork (Byrne, 2013). A quick signal of the existence of such arguments can prevent the dissemination of misinformation by identifying the location where such information was first discussed and providing alternative information. Organizations must consider whether they

are willing to use parts of their resources to act on these signals. Therefore, big data analyses need validation and this is often ensured through correlations with data that is already accepted as a legitimate foundation.

***Knowledge for development based on correlations.*** Both cases show that knowledge claims based on BD4D rest more on correlation than causation. BD4D often helps to answer "what" but usually falls short of explaining "why". The belief is that the detection of correlations is *in some cases* a better heuristic than searching for causal explanations. This propels BD4D applications in certain directions – i.e. towards cases where knowing what, but not why, is "good enough" (Mayer-Schönberger and Cukier, 2013). Statistics-based science is usually hypothesis-driven and thus designed to address specific questions about cause and effect, while BD4D is more exploratory and less focused on testing single relationships. This focus allows BD4D to include more variables into the analysis and explore surprising correlations.

The Kenya project is ultimately legitimized through a correlation between data on patterns of mobile use and official data on malaria prevalence. The project contains no data that can actually show *how* individual mobile users have caused malaria outbreaks in specific places. But the combination of the two data sets shows a correlation between movements and outbreaks. For instance, it can be shown that the return of Nairobi residents from vacations in infected areas correlates with the rise of malaria in the city. This correlation is then used to establish a theory about "returning residents" who bring home malaria to cities from risk regions. However, the actual transmission of the disease is not observed. We see a similar focus on correlations as a central aspect of justifying the use of BD4D data when turning to the case of using tweets to aid decision makers. A central element in the UN Global Pulse's justification for working with tweets as crisis signals is that rice prices approximated through sentiments in tweets do in fact correlate with official inflation rates. Again, the tweets do not

say much about why the inflation rate goes up or down, but – once validated through this correlation – it is a helpful piece of information in countries where statistics on inflation are lacking or produced slowly (UN Global Pulse and Crimson Hexagon, 2011).

## IMPLICATIONS – PROBLEMATIZING THE ROLE OF DATA IN INTERNATIONAL DEVELOPMENT

BD4D constitutes what Dean (2009: 32) terms a "problematization" which questions the role of data and established regimes of practices in international development. BD4D problematizes *what kind of data is used* in international development, especially the velocity of data. While traditional data-processing tools like household surveys can capture trends, BD4D rests on real-time data that enables swifter interventions. Both the *episteme* and *techne* of BD4D emphasize this type of problematization. The technical infrastructure (especially the sourcing software) enables fast-paced empirical sensitivity, and the resulting knowledge claims provide swift insights into problem areas. These claims are legitimized through visualization techniques that depict real-time data streams. Given that some areas of development work are criticized for slow response times, e.g. vaccination and disaster relief (Takeda and Helms, 2006), problematizing the velocity of data is central.

BD4D also problematizes *how data is collected and analyzed*. BD4D's *techne* rests on interfaces for sourcing large amounts of data and algorithms for synthesizing these, while the resulting *epistemes* are based on pattern-recognition rather than hypothesis-testing. Also, BD4D emphasizes the importance of granular empirical sensitivity – i.e. the ability to zoom in on details of a larger dataset (e.g. sub-groups, anomalies) – which is usually not captured through randomized sampling. This aligns some BD4D applications with recent calls for

disaggregating and contextualizing development data in order to offer more focused interventions (United Nations, 2013).

While BD4D questions the rationalities underlying established data-processing methods, a number of aspects of its emergence can be interrogated. Three aspects are particularly important. First, there is the challenge of detecting *relevant* anomalies. Many BD4D applications are based on the detection of anomalies, but it is not always easy to judge what is an unusually large or small amount of keywords, and this judgment may differ from one context to another. While BD4D may be able to detect anomalies from larger datasets, it cannot judge these anomalies in context.

Second, we must consider that many applications rest on data derived from people's own perceptions at a given moment in time (e.g. health symptoms). BD4D applications assume that these perceptions "correctly" reflect whatever is being analyzed – i.e. that a combination of keywords in a search query indicates that someone suffers from a disease. BD4D applications may create misleading results if the presence of flu-like symptoms is equated with the flu (Liu, 2010). Third, even though BD4D rests on larger data sets this does not imply that big data analyses produce better representations of the populations that development policies target. A variety of applications analyze social media interactions, but these constitute a sub-set of the overall population. As Boyd and Crawford (2012: 669) argue: "Twitter does not represent 'all people', and it is an error to assume 'people' and 'Twitter users' are synonymous; they are a very particular sub-set."

## CONCLUSION AND FUTURE RESEARCH

As UN Secretary-General Ban Ki-Moon remarked (United Nations, 2011): "our traditional 20th century tools for tracking international development cannot keep up." This article responds to this situation by discussing the emergence of BD4D as an ever more widely recognized regime of data processing in international development. Our study has no explicit evaluative dimension. We neither argue that BD4D is a better (or the best) way to collect and analyze development data, nor do we claim that traditional statistics at both global and local levels will be replaced. Our approach seeks to unpack how big data conditions international development in novel ways, and to conceptualize these in terms of the focus of governmentality on visibilities, techniques and rationalities.

While our discussion emphasizes a number of ways in which development problems are approached through big data analyses, many issues require further attention. An important focus in governmentality approaches concerns the effects of governance regimes on human subjects and their identity formation. The Foucauldian interest in "subject formation" (Foucault, 1982) gives attention to the identities formed in and through regimes of practices and explores how subjects internalize ways of acting and thinking that condition and shape their conduct. BD4D may lead to significant re-configurations of professional identities and the identities of subjects on the ground, such as aid recipients. BD4D requires data-savvy managers and analysts whose professional identities differ from those of bureaucrats (Hilbert, 2013). The further institutionalization of BD4D will most likely involve clashes between different professional and organizational identities, such as what it means to be a good researcher or an innovative, data-driven organization. Future research needs to explore how such identities are shaped, how identity formation interacts with BD4D's *techne* and *episteme*, and also how professional and organizational identities can be aligned.

BD4D also has ramifications for identity formation and human conduct on the ground. Applying big data to address development challenges presupposes a specific subject to be

governed – primarily the young, media-savvy and connected. This makes BD4D itself a "dividing practice" (Dean, 2009: 156) that categorizes and assumes a certain type of conduct from the subjects to be governed. BD4D draws a line between subjects whose digital traces inform the analysis of development problems and those who are either unwilling or unable to contribute. Future research needs to debate the extent to which this limits the applicability (e.g. by excluding older people or those living in rural areas).

Future research should also examine how BD4D produces intersections and clashes between different forms of knowledge production. For instance, the effects of a growing reliance on algorithmic intelligence on well-established forms of knowledge, such as those stemming from statistical agencies, deserve scrutiny. How can these forms of knowledge meaningfully complement each other when crafting and focusing governance efforts? How does the rationalization of knowledge impact the ways individuals and organizations are made accountable for the effects of policies and governance initiatives? We believe such questions are important and timely since they show how the current disconnect between big data analyses, traditional statistics and other forms of knowledge production can be overcome.

## REFERENCES


Althouse, B. M., Y. Y. Ng, and D. Cummings. 2011. Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis* 5(8): e1258.

Amoore, L., and V. Piotukh. 2015. Life beyond big data: Governing with little analytics. *Economy and Society* 44: 341-366.

Anderson, C. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired* 16: 108-109.

Bayir, M.A., M. Demirbas, M., and N. Eagle. 2009. Discovering spatiotemporal mobility profiles of cellphone users, *World of Wireless, Mobile and Multimedia Networks & Workshops*, IEEE, pp. 1-9.

Billari, F., F. D'Amuri, and J. Marcucci. 2013. Forecasting births using Google, Paper presented at *Annual Meeting of the Population Association of America*, New Orleans, LA.

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

Bowker, G. C., and S. L. Star. 2000. *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Boyd, D. and K. Crawford. 2012. Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15: 662-679.

Byrne, C. 2013. *How the UN's new data lab in Indonesia uses Twitter to preempt disaster.* Available at: http://www.fastcolabs.com/3007178/open-company/how-uns-new-data-lab-indonesia-uses-twitter-preempt-disaster.

Chen, H., R. H. L. Chiang and V. C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly* 36: 1165-1188.

Chinn, M. D., and R. W. Fairlie. 2010. ICT use in the developing world: An analysis of differences in computer and internet penetration. *Review of International Economics* 18: 153-167.

Chunara, R., J.R. Andrews, and J.S. Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene* 86(1): 39–45.

De Choudhury, M., M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. 2010. Automatic construction of travel itineraries using social breadcrumbs, *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, ACM: 35-44.

Dean, M. 1996. Putting the technological into government. *History of the Human Sciences* 9: 47–68.

Dean, M. 2009. *Governmentality: Power and rule in modern society* (2nd edition). London et al: SAGE.

Deaton, A. 2000. *The analysis of household surveys*. Baltimore: Johns Hopkins University Press.

Deville, P., C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondel, and A.J. Tatem. 2014. Dynamic population mapping using mobile phone data, *Proceedings of the National Academy of Sciences*, 111(45): 15888-15893.

Ettredge, M., J. Gerdes, and G. Karuga. 2005. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM* (Association for Computing Machinery), 48(11): 87–92.

Ferrari, L. and M. Mamei. 2011. Discovering daily routines from Google latitude with topic models, *Pervasive Computing and Communications Workshops* (PERCOM Workshops), IEEE: 432-437.

Ferrari, L., A. Rosi, M. Mamei, and F. Zambonelli. 2011. Extracting urban patterns from location-based social networks, *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, ACM: 9-16.

Foucault, M. 1978. *Discipline and punish: The birth of the prison*. New York: Vintage.

Foucault, M. 1982. The subject and power. *Critical Inquiry* 8: 777-795.

Foucault, M. 1991a. Governmentality. In *The Foucault effect: Studies in governmentality*. eds. G. Burchell, C. Gordon, and D. Murphy, pp. 87-104. Harvester.

Foucault, M. 1991b. Questions of method. In *The Foucault effect: Studies in governmentality*. eds. G. Burchell, C. Gordon, and D. Murphy, pp. 73-86. Harvester.

Frías-Martínez, V., and Virseda, J. 2013. Cell phone analytics: Scaling human behavior studies into the millions. *Information Technologies & International Development* 9: 35–50.

Frías-Martínez, E., G. Williamson, and V. Frías-Martínez. 2011. *An agent-based model of epidemic spread using human mobility and social network information*. In Privacy, security, risk and trust (passat), IEEE Third International Conference on Social Computing (socialcom): pp. 57–64.

Gaver, W. W. 1991. Technology affordances. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching Through Technology*: 79–84.

Gillespie, T. 2014. The relevance of algorithms. In *Media technologies: Essays on communication, materiality, and society*, eds. Gillespie, T., P. Boczkowski, and K. Foot, pp. 167-194. MIT Press.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–1014.

Gitelman, L. 2013. *Big data is an oxymoron.* Cambridge, MA: MIT Press.

Gouldson, A. and J. Bebbington. 2007. Corporations and the governance of environmental risk. *Environment and Planning C: Government and Policy 25*: 4–20.

Greiman, V. 2011. *Guide on international development: Public service careers and opportunities*, Available at: http://hls.harvard.edu/content/uploads/2008/07/developmentguidefinal.pdf?redir=1

GSMA. 2012. *Sub-Saharan Africa mobile observatory 2012*. London: GSMA.

Hilbert, M. 2013. *Big Data for development: From information- to knowledge societies*. Available at: http://ssrn.com/abstract=2205145.

Howard, P. N., and N. Mazaheri. 2009. Telecommunications reform, Internet use and mobile phone adoption in the developing world. *World Development* 37: 1159–1169.

Hubbard, D. W. 2011. *Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities*. Hoboken, NJ: John Wiley.

Ilcan, S., and L. Phillips. 2010. Developmentalities and calculative practices: The Millennium Development Goals. *Antipode* 42: 844–874.

International Telecommunication Union (ITU). 2012. *Trends in telecommunication reform in 2012: Smart regulation for a broadband world*. Geneva: ITU.

King, G., and E. N. Powell. 2008. *How not to lie without statistics.* Working Paper, Harvard University Institute for Qualitative Social Science Research, Cambridge, MA.

Kohut, A., R. Wike, J. M. Horowitz, K. Simmons, J. Poushter, C. Barker, J. Bell, and E.M. Gross. 2011. *Global digital communications: Texting, social networking popular worldwide*. Pew Global Attitudes Project: Pew Research Center, Washington, DC, USA.

Laney, D. 2001. *3D data management: Controlling data volume, velocity and variety*. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Latour, B., P. Jensen, T. Venturini, S. Grauwin and D. Boullier. 2012. 'The whole is always smaller than its parts': A digital test of Gabriel Tardes' monads. *The British Journal of Sociology*, *63*: 590-615.

Letouzé, E. 2012. *Big data for development: Opportunities and challenges*. New York: United Nations Global Pulse.

Liu, B. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing,* eds. N. Indurkhya N and F.J. Damerau, pp. 627-665, Taylor & Francis.

Lohr, S. 2013. Searching big data for 'digital smoke signals'. *New York Times,* August 7. Available at: http://www.nytimes.com/2013/08/08/technology/development-groups-tap-big-data-to-direct-humanitarian-aid.html.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big data: A Revolution that will transform how we live, work and think***.** London: John Murray.

McAfee, A., and E. Brynjolfsson. 2012. Big data: The management revolution. *Harvard Business Review* 90: 60–68.

Merry, S.E. 2011. Measuring the world: Indicators, human rights, and global governance, *Current Anthropology*, 52: 83-95.

Methmann, C. 2011. The sky is the limit: Global warming as global governmentality. *European Journal of International Relations* 19: 69–91.

Miller, P., and N. Rose. 1990. Governing economic life. *Economy and Society* 19: 1-31.

Mocanu, D., A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignan. 2013. The Twitter of babel: Mapping world languages through microblogging platforms. *PLoS ONE,* 8(4): e61981.

Morales, A-J., D. Pastor-Escuredo, Y. Torres, V. Frías-Martínez, E. Frías-Martínez, N. Oliver, A. Rutherford, T. Logar, R. Clausen-Nielsen, O. de Backer, and M.A. Luengo-Oroz. 2015. *Studying Human Behavior through the Lens of Mobile Phones during Floods*, International Conference on the Analysis of Mobile Data, NetMob.

Moumni, B., V. Frías-Martínez, and E. Frías-Martínez. 2013. Characterizing social response to urban earthquakes using cell-phone network data: The 2012 Oaxaca earthquake. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing*: 1199–1208 New York, NY: ACM.

Murray, L. T. 2007. *The will to improve: Governmentality, development, and the practice of politics*. Durham, NC: Duke University Press.

Neuhaus, F., and T. Webmoor. 2012. Agile ethics for massified research and visualization. *Information, Communication & Society* 15: 43-65.

Piotrowski, J. 2013. *UN initiative mines big data to direct development*. Available at: http://www.scidev.net/global/data/news/un-initiative-mines-big-data-to-direct-development.html.

Reis, B.Y. and J.S. Brownstein. 2010. Measuring the impact of health policies using internet search patterns: The case of abortion, *BMC Public Health*, 10: 514.

Ritterman, J., M. Osborne, and E. Klein. 2009. Using prediction markets and Twitter to predict a swine flu pandemic, *1st International Workshop on Mining Social Media*.

Rogers, R. 2013. Debanalizing twitter: The transformation of an object of study. Proceedings of the *5th Annual ACM Web Science Conference,* Paris*:* 356-365.

Samman, E. 2013. *Using household surveys to start a data revolution and tackle social inequality*. Available at: http://www.theguardian.com/global-development-professionals-network/2013/jun/10/mdgs-household-surveys-data-revolution.

Sumner, A., and M. A. Tribe. 2008. *International development studies: Theories and methods in research and practice,* London: Sage.

Takeda M.B. and M.M. Helms. 2006. "Bureaucracy, meet catastrophe:" Analysis of the tsunami disaster relief efforts and their implications for global emergency governance. *International Journal of Public Sector Management*, 19: 204-217.

The World Bank. 2004. *Monitoring and evaluation: Some tools, methods and approaches.* Washington D.C.: The World Bank.

UN Global Pulse and Crimson Hexagon. 2011. *Twitter and perceptions of crisis-related stress: Methodological white paper*. New York, NY: United Nations.

United Nations. 2005. *Household sample surveys in developing and transition countries*. Department of Economic and Social Affairs (ST/ESA/STAT/SER.F/96). New York, NY: United Nations.

United Nations. 2011. *Secretary-General's remarks at General Assembly briefing on the Global Pulse initiative.* Available at: http://www.un.org/sg/statements/?nid=5668.

United Nations. 2013. *A new global partnership: Eradicate poverty and transform economies through sustainable development (The report of the high-level panel of eminent persons on the post-2015 Development agenda).* New York, NY: United Nations.

Wang, X., M.S. Gerber, and D.E. Brown. 2012. Automatic crime prediction using events extracted from Twitter posts. In *Social Computing, Behavioral - Cultural Modeling and Prediction,* eds. S. J. Yang, A. M. Greenberg, and M. Endsley , pp. 231–238. Springer.

Wesolowski, A., N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. 2012. Quantifying the impact of human mobility on malaria. *Science*, 338(6104): 267-270.

Figure 1: Visualization of human movement in Kenya through cell phone data (Source: Wesolowski et al., 2013)

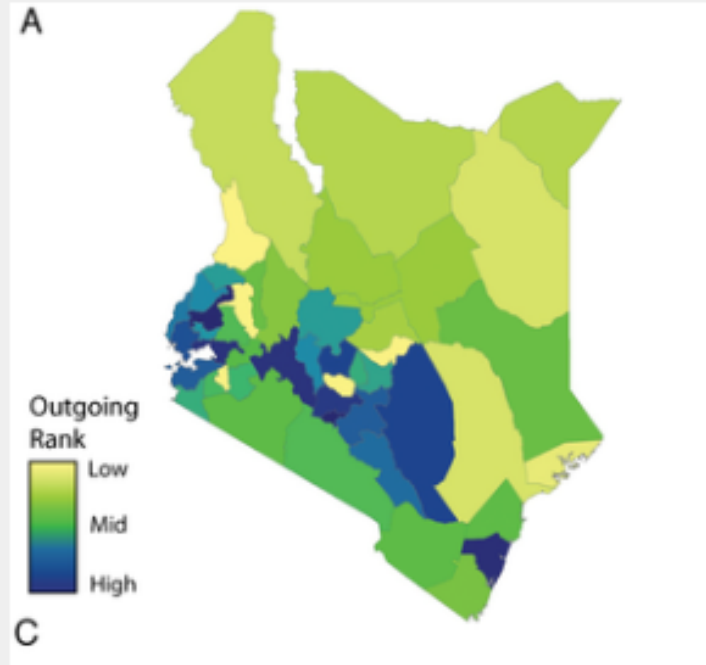

Figure 2: Twitter-based monitor to visualize economic crisis-signals in Indonesia (Source: UN Global Pulse and Crimson Hexagon, 2011)