Big Data as Governmentality – How Large-Scale Data Reshapes International Development

Abstract:

This paper conceptualizes how large-scale data and algorithms condition and reshape knowledge production when addressing international development challenges. The concept of governmentality and four dimensions of an analytics of government are proposed as a theoretical framework to examine how big data is constituted as an aspiration to improve the data and knowledge underpinning development efforts. Based on this framework, we argue that big data's impact on how relevant problems are governed is enabled by (1) new techniques of visualizing development issues, (2) linking aspects of the international development agenda to algorithms that synthesize large-scale data, (3) novel ways of rationalizing knowledge claims that underlie development policies, and (4) shifts in professional and organizational identities of those concerned with producing and processing data for development. Our discussion shows that big data problematizes selected aspects of traditional ways to collect and analyze data for development (e.g. via household surveys). We also demonstrate that using big data analyses to address development challenges raises a number of questions that can deteriorate its impact.

Keywords: big data; international development; governmentality; algorithms; knowledge production

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) 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 (Boyd and Crawford, 2012).

Research has mostly discussed big data as an opportunity for firms to increase their market share and competitiveness. McAfee and Brynjolfsson (2012), for instance, argue that big data helps businesses gather real-time data about customer behavior and develop targeted advertising and improved decision-making, while Bughin, Livingston and Marwaha (2011) claim that big data analytics need to be carefully aligned with a firm's overall strategic direction in order to maximize its potential. Surprisingly little analysis is directed towards understanding the practices underlying other uses of big data (for an exception see Hilbert, 2013).

However, policymakers are realizing the potential of big data to produce actionable information that can be used to improve development (e.g. related to poverty reduction) – e.g. by identifying needs, providing services, and predicting crises (World Economic Forum, 2012; Ginsberg et al., 2009). The aim of this paper is to analyze how big data analyses condition and contribute to international development. We are interested in the mechanisms by which big data renders certain areas of international development governable. More precisely, we aim to explore the practices and rationales of governance that allow aspirations of reform, such as big data for development (hereafter BD4D), to be constituted.

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 governmentality lens to explore the rationales, practices and power structures underlying international development (Methmann, 2011; Murray Li, 2007), we know very little about how changes in sourcing, processing and communicating relevant 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 make it possible for big data to have an effect on the way knowledge for international development is produced. We structure the discussion around Dean's (2009) four 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, the forms of knowledge (*episteme*) attached to certain regimes, and the forms of identity formation that belong to them. Together these four dimensions provide a framework to examine the specific conditions under which BD4D emerges as a way of processing data to address development challenges. Using these dimensions in our analysis, we argue that the effective 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" and instead relying on data provided by a distributed set of private companies with

proprietary algorithms (e.g. Twitter), (3) the acceptance of new epistemic foundations for governing development problems (e.g. when creating policies), and (4) the acknowledgment that working with big data challenges professional and organizational identities (e.g. when "traditional" development analysts need to turn into data-savvy managers). We suggest that, taken together, these dimensions problematize selected aspects of established data processing practices used in the field of development (e.g. household surveys).

Our analysis does *not* argue that big data replaces traditional ways to handle data in international development, nor do we argue that BD4D is without problems. Rather we suggest that BD4D complements established ways of handling development data, because it opens new perspectives on problems and adds depth and speed to the analysis. With this in mind, 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, yet subtle, form of power shaping how problems and opportunities are framed and acted upon. Second, we contribute to discussions on how to govern the 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 governmentality effects of working with large-scale data remain under-explored.

BIG DATA FOR DEVELOPMENT

Big Data

The amount of data produced by human activities increases rapidly and currently needs to be counted in zettabytes (trillions of gigabytes). In 2012, the world's data amounted to 2.8 zettabytes, and this is expected to double every second year and reach 40 zettabytes by 2020 (Gantz and Reinsel, 2012). Such digital traces span from travel patterns captured by GPS

devices, over Internet traffic and searches stored online to messages on social media. The term ”big data” is used to describe these growing amounts of data and their uses, which require new forms of data storage, analysis and visualization (Chen et al., 2012), and offer new possibilities for measurement, prediction and governance.

The current focus on big data revolves around the possibilities offered by widespread “datafication” (Mayer-Schönberger and Cukier, 2012: 73), whereby multiple elements of social life are quantified and hence take the shape of digital data, such as when our travels and other movements become traces left by GPS devices, and friendships take the shape of “likes” on Facebook. The resulting masses of data, coupled with the emergence of advanced data mining technologies and visualization techniques, constitute the foundation of big data. In trying to make sense of this phenomenon, some accounts focus on the growth of digital traces, such as the growing *volume*, *velocity* and *variety* of data (Laney, 2001), or outline how algorithmic analyses are able to integrate such large-scale, fast-moving and disparate forms of data based on correlations (Boyd and Crawford, 2012).

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 conceptualize big data as a phenomenon taking shape at the intersection of the growing velocity, variety and volume of datafication processes (Laney, 2001; Mayer-Schönberger and Cukier, 2013), and algorithmic developments allowing for mining, correlation and visualization of digital traces from disparate sources (Boyd and Crawford, 2012; 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 for purposes of prediction, measurement and governance*.

Big Data for Development (BD4D)

Discussing big data as a way to improve international development rests on the increasing awareness that the production of large amounts of data is not 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, transmit medical information, and buy/sell goods. Data generated through such usage can produce information that is relevant when addressing development challenges. Many developing and emerging countries have also witnessed swiftly rising social media growth rates (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), it is fair to say that there is potential in using big data to address selected development problems. Recently, a number of organizations have started to unlock this potential (see Table 1).

==========

Insert Table 1

==========

One way to use big data to create actionable information for development is by tracking words (Hilbert, 2013). Facebook posts, tweets, Google searches, and blogs generate a lot of

relevant real-time data. Analyzing this data can yield indicators that predict trends (e.g. related to the outbreak of diseases), often with the same statistical significance as traditional data collection tools (Ginsberg et al., 2009). For instance, the UN Global Pulse initiative analyzed large amounts of tweets commenting on the price of rice in Indonesia. The analysis showed that the quantity of tweets on the topic followed the official inflation for the food basket in the country, indicating that social media data can be used as a predictor of price trends on local markets (UN Global Pulse and Crimson Hexagon, 2011). Such applications of big data reduce the time lag between the emergence of a trend and responses by governments and aid agencies, as the data is sourced and analyzed in real-time (see also UNICEF, 2013).

Social media data can also be used in combination with more traditional data from mobile devices (e.g. text messages). The crowdsourcing platform Ushahidi (meaning "testimony" in Swahili) was designed to turn data from different channels into real-time crisis maps that can assist humanitarian relief efforts. Ushahidi launched a crisis map within four days after the Haitian earthquake in 2010. This crowdmap visualized information provided by those who were affected and by those who were involved in humanitarian efforts (e.g. by locating broken bridges/roads). Other applications of big data for development have emerged by tracking location-based data (e.g. GPS data from mobile devices) and nature-related data (e.g. weather). All of these applications rely on automated algorithms to enable fast data analysis. For instance, following tweets about rice prices requires building an algorithm that follows rules for separating out non-topical tweets where "rice" has wrong connotations. Similarly, crisis-maps in real-time require algorithms to sort in the temporality of information being launched onto the map.

So far, international organizations and national governments primarily use survey-based data to create, monitor and evaluate development policies (Ginsberg et al., 2009; United Nations, 2013). The UN and the World Bank frequently use household surveys for collecting

information on populations in developing and emerging economies (The World Bank, 2004; United Nations, 2005). These provide data on a variety of topics, ranging from poverty to healthcare and education. 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 head of the household, 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 available with some delay (Deaton, 2000). It is for this reason that BD4D can *supplement* survey-based data in some selected contexts, depending on what kind of data is available, by whom, on what topics, and in which regions.

GOVERNMENTALITY: EXPLORING THE CONDITIONS FOR GOVERNING ACTION

The concept of governmentality builds on and extends Michel 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 political meaning of the word. Government can be understood as a form of power (i.e. the power to shape human conduct). The governmentality literature offers an analytical vocabulary that is useful if we want to understand and conceptualize the *workings of government*. Governmentality is primarily concerned with "how" questions (Dean, 2009: 39). It explores the conditions of governing ("How do we govern?" and "How are we governed?") by focusing on what is necessary to make a particular regime of practices work.

The main strength of a governmentality perspective is that it views governance as a complex of operations, calculations and reasonings that condition particular attempts to shape conduct. The concept of governmentality views power not as something that is allocated by means of structural properties (e.g. state policies). Rather power is exercised "at a distance" (Miller and Rose, 1990) and is constituted through governmental practices and their underlying technologies and rationalities.

Dimensions of an Analytics of Government

One extension of Foucault's thoughts 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 reflect stable and organized patterns by which we do things (e.g. collecting data). Such regimes include taken-for-granted aspects of technologies, mentalities, agencies, and visibilities (Dean, 2009: 37) and they define the aim of relevant practices and their objects. Practically speaking, we find a lot of interrelated regimes of practices within organizations. Often, we refer to such regimes as systems (e.g. data processing system), as they cannot be clearly ascribed to single individuals or groups (Spence and Rinaldi, 2012). Our analysis is concerned with big data as a regime of practices for collecting and analyzing data relevant to international development.

Dean (1995, 1996, 2009) has introduced the following four dimensions of an analytics of government that offer a framework to critically analyze the conditions under which regimes of practices operate. The four dimensions are co-present within regimes of practices and presuppose one another.

Fields of visibility. What is rendered visible and invisible shapes processes of governance to a significant degree. This first dimension is concerned with the field of

visibility that characterizes a given regime of practices (Dean, 2009: 41). There are many ways to make objects of government visible: flow charts, data tables, maps, and organograms are just some examples. Ways of visualizing simultaneously define objects of government, as they highlight some characteristics while hiding others. This is particularly true when considering that a great deal of knowledge in and between organizations assumes the form of visual representations (e.g. through PowerPoint, Gabriel, 2008). Visualizations may even obscure certain objects completely (e.g. when a flow chart drops a process). Dean stresses that a field of visibility involved in a regime of practice maps who and what is to be governed, what governance problems are to be addressed, and with what objectives in mind.

The techne of government. 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 technologies to refer to mechanisms like rankings that normalize behavior (Sauder and Espeland, 2009), compliance mechanisms like codes of conduct, and techniques of education (Spence and Rinaldi, 2012). 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.

The episteme of government. *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 (Miller and Rose, 1990). *Episteme* is hence understood as revealing assumptions about how knowledge

involved in practices of governing is justified and privileged, such as the truth claims that big data gives rise to.

Identity Formation. The last dimension emphasizes the forms of identity that are formed in and through regimes of practices, and which influence the way government functions (Dean, 2009: 43). This dimension puts "subject formation" at the center of analysis (Foucault, 1982: 777) and explores how subjects internalize certain technologies and rationalities and translate these into conduct. When thinking about how subjects are formed in and through regimes of practices we can focus on those who exercise authority or on those who are governed, examining how a given regime of practices conditions their conduct and presupposes certain types of persons. Dean (1990) emphasizes that this dimension focuses on processes of identification. The question is what drives and limits the enactment of particular identities and how does this affect the government of conduct?

ANALYZING BIG DATA FOR DEVELOPMENT AS GOVERNMENTALITY

Our analysis treats BD4D as an emerging regime of data processing in organizations that support international development. An analytics of government reveals the conditions of the emergence of this regime of practices by examining how it gives rise to and depends upon new forms of knowledge, relies on algorithmic intelligence as a technology, challenges professional identities, and promotes new ways of visualizing development problems.

The Field of Visibility Characterizing BD4D

The operation of BD4D as a regime of practices produces certain forms of visibility, picturing what is being governed. The forms of visibility connected to BD4D range from

traditional ways of depicting development data (e.g. trend charts) to new visualization techniques (e.g. word clouds). Together these forms tap into digital traces of everyday practices and repurpose them to visualize aspects of the field of international development and hence make it governable. The UN Global Pulse has used crisis monitors to visualize how new types of data can help to understand in what ways people are impacted by food prices (see Figure 1). These monitors rely on algorithmic recognition of patterns in the semantic content of tweets as a foundation for transforming real-time data feeds into word clouds, semantic clusters, and color-coded topics that enable users to detect weak signals (e.g. when non-affordability impacts living conditions). Real-time maps are another way of visualizing BD4D. Figure 2 shows an Ushahidi crisis map of Haiti's capital Port-au-Prince after the 2010 earthquake. The map combines different types of data, such as location and infrastructure data, within a single visualization.

==========

Insert Figure 1

==========

==========

Insert Figure 2

==========

Novel visualization techniques better capture the characteristics of big data and hence make development problems governable in different ways. For instance, depicting real-time location-based data increases the response time of humanitarian relief efforts (Meier, 2012). But 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 (Reid and

Frankel, 2008). The "how" of visualization is important because choices concerning the logics of categorization beneath the visual representations are controversial. In the process of building the Twitter visualization in Figure 1, for instance, it was an explicit priority for the UN to give topical experts a central role in "training" the algorithm to return a visualization with categories that were recognizable and sensible within the organization (UN Global Pulse and Crimson Hexagon, 2011). Such choices concerning the balance between algorithmic intelligence and human expertise is at the core of producing actionable visualizations for developmental purposes.

Visualizations offer new perspectives on information contained in a dataset that would otherwise be hard to grasp. The crisis monitor depicted in Figure 1 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. Visualizations that allow for gradual shifts between levels of analysis offer new understandings of how, for instance, rising food prices impact people.

The Technical Aspects of Governing Through BD4D

Governing international development through BD4D is shaped by the *affordances* of computational technologies. The affordance of a technology designates the way in which it supports or limits certain uses (Gaver, 1991), and a focus on technologies and their affordances invites us to unpack how governance is influenced by the nature of the data and computational technologies involved. Data used by BD4D applications is often formatted in particular ways in terms of size and the amount of meta-data involved. For instance, 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 (@). These formats produce data

with specific affordances, conditioning the kinds of knowledge that can be produced. When an organization like the UN repurposes Twitter's data, it chooses to work with short, condensed messages from a specific communicative culture that ultimately 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.

To understand in what ways technologies condition the way BD4D shapes international development, we need to consider the means by which digital traces are sourced, aggregated and visualized. We distinguish between two types of technologies that are central to BD4D: (1) technologies that can harness and quantify real-time behavioral data *(sourcing software)* and (2) technologies that provide the formal instructions needed for a computer to transform inputs from the sourcing software into visual outputs *(synthesizing algorithms)*.

Sourcing software. The relevance of sourcing software can be illustrated by returning to the UN Global Pulse crisis monitor (see Figure 1). The visual cues in the monitor are dependent on a chain of sourcing software. A central element in this chain is the "pipeline" through which Twitter offers external partners an opportunity to interact with their database. This pipeline is called an Application Program Interface (API) and it shapes how streams of tweets can be sourced and repurposed. Hence, when trying to understand how Twitter-based BD4D applications afford certain uses, one must understand the characteristics of the relevant API. The standard output of Twitter's public API is a random sample consisting of about one percent of all the tweets. Also, the provenance (i.e. lineage, origins and sources) and granularity (i.e. level of detail and inclusiveness) of the data are often difficult to establish, as such aggregates are hard to trace back to individual sources (Neuhaus and Webmoor, 2012). In order to get a more comprehensive and transparent sample, organizations like the UN have to either negotiate other terms with Twitter or team up with an approved reseller of Twitter's data such as Crimson Hexagon, which is the UN's data partner in the Twitter study.

Synthesizing algorithms. The output of a chain of sourcing software is a set of quantitative data that can be processed as signals of human behavior. For instance, the data sourced from Twitter's API comes in excel-like spreadsheets that are readable for a computer (Neuhaus and Webmoor, 2012). Synthesizing algorithms do this further processing based on logical rules that give instructions as to how data is processed and organized. The algorithms that turn quantified data into visual representations are central when using big data in international development. In the context of the UN Global Pulse's crisis monitors, algorithms are used to compute semantic distances between words in tweets on the basis of theories of natural language processing and network characteristics, and turn these into semantic networks (see middle of Figure 1).

Synthesizing algorithms partly displace the power of theoretical interpretation from topical experts (who draw on past experience) to computer scientists (who rely on machine intelligence). Anderson (2008) referred to this development as "the end of theory" – i.e. automated algorithms can potentially generate more useful insights than traditional research strategies (e.g. hypothesis-testing). Although it has already been emphasized that BD4D applications also rely on experts (e.g. to decide whether data categories rest on the frequency or mere presence of keywords; King and Powell, 2008), it is clear that BD4D plays up the role 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.

The Forms of Knowledge Attached to BD4D

New forms of knowledge arise from and inform BD4D. Big data changes how knowledge is rationalized and hence creates a different ground upon which to evaluate

"truth". This rationalization differs from more established forms of knowledge production in three ways.

Knowledge for development based on masses of data. Household surveys rely on rather small amounts of sampled data, because collection is time and resource intensive (Deaton, 2000), and define up front what kind of data is needed and how it is to be used. By contrast, BD4D applications rest on an inductive analysis of large amounts of unstructured data. On the one hand, this allows for more explorative analyses when trying to understand development problems, as it shows new dimensions of existing problems or even helps to unravel new problems altogether. On the other hand, this makes knowledge claims more dependent on how algorithms work (and not work). Algorithms often take past experiences as a predictor of the future. This can limit their potential to cope with the constantly changing environments in which some development problems are embedded, even when considering that new computer-based simulations combine data from the past with future scenarios (Hilbert, 2013).

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 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 Google searches or tweets, instead of a few hundred surveys, makes the resulting knowledge *appear* legitimate, even though the data underlying BD4D is also constrained in numerous ways, and often just a by-product of peoples' engagement with digital devices. Combining different forms of knowledge further enhances legitimacy. For instance, Google Dengue Trends shows how Google's query-based estimates match the official statistics by the Brazilian Ministry of Health (see Figure 3). This matching increases legitimacy, as it connects the results of big data analysis to a form of knowledge more familiar to policymakers.

==========

Insert Figure 3

==========

Knowledge for development based on messy but real-time data. When creating knowledge for development by increasing the scale of samples, we have to move beyond clean, carefully designed data and accept some messiness (Cukier and Mayer-Schönberger, 2013). While surveys operate on the assumption that data needs to be as correct as possible (since the sample size is limited), BD4D can afford some inaccuracies (e.g. the inclusion of *some* irrelevant Google searches) 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. Also, 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 in crisis intervention.

Much of the authority given to knowledge resulting from BD4D applications rests on the combination of large-scale and small-scale data analysis. BD4D is a good way to identify "digital smoke signals" (Lohr, 2013) – i.e. pulses, anomalies, and trends that survey-based methods rarely capture. These weak signals can then be used to further investigate a problem. For instance, the UN Global Pulse's analysis of Twitter data in Indonesia revealed that people were saying that vaccines were not halal as they contain pork (Byrne, 2013). BD4D can prevent the dissemination of such misinformation, by identifying the location where such information was first discussed and providing alternative information. The immediate availability of results through BD4D applications strengthens the authority of knowledge

claims and offers the possibility to adjust policies much quicker, allowing those in charge of development work to see what is (not) working (Piotrowski, 2013).

Knowledge for development based on correlations. Knowledge claims based on BD4D rest more on correlation than causation. Google Flu Trends can predict to some degree seasonal influenza outbreaks, but it cannot tell us *why* outbreaks in certain locations occur. 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. The fact that the price of rice in Indonesia (when approximated through tweets) correlates with the official inflation rate does not say much about why the inflation rate goes up or down, but it is a helpful piece of information in a country where reliable statistics on inflation are lacking (UN Global Pulse and Crimson Hexagon, 2011). This directs BD4D applications into a certain direction – i.e. towards cases where knowing what – but not why – is "good enough" (Cukier and Mayer Schönberger, 2013).

Identity Formation in the Context of BD4D

Professional Identities in Organizations. BD4D initiatives operate in diverse organizational contexts. Some initiatives are designed as non-profit entities (e.g. Ushahidi), while others are embedded in international organizations (e.g. UN Global Pulse). The formation of identities is particularly interesting in the latter context. The reliance on big data as a supplement to more traditional forms of knowledge production means that an initiative such as UN Global Pulse needs to establish itself as a legitimate source of information *within* the UN system. Given that the UN is often characterized as a bureaucracy, lacking innovation and flexible structures (Jaeger, 2010), a shift towards BD4D challenges established identity patterns. In a recent interview with the director of UN Global Pulse, Robert Kirkpatrick, this

initiative was described as an "Internet start up", "an exercise in entrepreneurship", and an attempt to "track unemployment and disease as if it were a brand" (Lohr, 2013: 1-2). All of these attributes do not seem very characteristic of the UN.

BD4D initiatives in these organizations cannot operate in isolation, as they depend on other organizational parts that have the relevant expertise to translate their results into action. Also, the success of BD4D applications depends on having data-savvy managers and analysts who often uphold different professional identities than bureaucrats (Hilbert, 2013). Governing through BD4D requires thinking about how newly emerging patterns of professional identity can be integrated in and aligned with those organizations that host relevant initiatives.

The identity of the governed subject. While many of the digital traces that feed into BD4D are naturally occurring bi-products of online activities, applying big data to address development challenges presupposes a specific subject to be governed – primarily the young, media-savvy and connected. While there is nothing wrong with presupposing this ideal-type identity, it shows that applications assume a certain type of conduct from those who are to be governed. This makes BD4D itself a "dividing practice" (Dean, 2009: 156) – i.e. a practice that categorizes the subjects to be governed. Even though not an intended outcome, BD4D draws a line between subjects whose digital traces inform the analysis of relevant development problems and those who are either unwilling or unable to contribute. While this divide conditions the emergence of BD4D, it also limits its applicability. For instance, applications are likely to exclude older people or those living in rural areas without a well-developed technical infrastructure (Norris, 2001). This, in turn, limits the demographic and geographic scope of applying BD4D. As Chan et al. (2011) have argued, Google Dengue Trends faces challenges in rural areas where sufficient search volume could not be reached.

IMPLICATIONS – PROBLEMATIZING THE ROLE OF DATA IN INTERNATIONAL DEVELOPMENT

We have discussed how big data makes selected areas of international development amenable to governance. The analysis demonstrates how new rationalities inscribe themselves into regimes of practices of data processing. This is *not* to say that BD4D replaces established data processing practices. It rather offers an alternative way of thinking about the role of data in international development, asking us to reflect on the limits of established techniques for data collection and analysis. Thus, BD4D constitutes what Dean (2009: 32) terms a "problematization" that questions the role of data and established regimes of practices in international development, and outlines alternatives.

First, BD4D problematizes *what kind of data is used* in international development, especially the velocity of data. Moving from survey-based data to big data reduces the time lag between the start of a trend and the response by governments and other authorities. While traditional data processing tools like household surveys can capture these trends, BD4D rests on real-time data that enables quicker 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, while the resulting knowledge claims are at least partly built upon providing swifter insights into problem areas. The discussion of visibility shows how these claims are legitimized through visualization techniques that depict real-time data streams. Given that some areas of development work are constantly criticized for slow response times (e.g. vaccination and disaster relief; Takeda and Helms, 2006), problematizing the velocity of data seems important and timely.

Second, BD4D 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. The discussion of identity formation showed that this way of treating data presupposes a subject who datafies certain aspects of life. All of this problematizes sampling-based approaches to producing development data. 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 are usually not captured through randomized sampling. Google Flu Trends, for instance, enhances the analysis of the spread of influenza by providing city-level data (Ginsberg et al., 2009), while surveys focus more an aggregated regional data. This aligns some BD4D applications with recent calls for disaggregating and contextualizing development data in order to offer more focused interventions (United Nations, 2013).

Finally, BD4D problematizes *data-related capacities and skills* in organizations supporting development work. Our analysis of BD4D's *techne* showed the importance of new software tools, and the discussion of identity formation revealed the need to align the skills required for handling these tools with professional and organizational identities. Such skills are not only needed to operate BD4D applications, but also to review how BD4D analyses are performed, thereby ensuring accountability and transparency, which is vital when challenging established analytical techniques. Without introducing new data experts (primarily computer scientists and mathematicians; Nelson, 2008), there is a risk that the algorithms and datasets behind analyses will become black boxes (Mayer-Schönberger and Cukier, 2013).

Problematizing BD4D

While BD4D questions the rationalities underlying established data processing methods, a number of aspects of its own 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. However, what is considered as e.g. an

unusually high/low amount of keywords is not always easy to judge and may differ from one context to another. While BD4D may be able to detect anomalies out of larger datasets, it cannot judge these anomalies *in context* (Lazer et al., 2014). Returning to one of the examples discussed above, the degree of abnormality of the price of rice is likely to differ according to the geographic and cultural context within a given country.

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 has the flu. However, perceptions are not objective facts, but subjective expressions of individuals' perspectives. Hence, BD4D applications can 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 population 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." BD4D's emergence as a complementary regime of practices for addressing development challenges will depend on acknowledging such limitations.

CONCLUSION AND FUTURE RESEARCH

BD4D reconfigures the role of data within international development and delineates ways in which large-scale data can supplement established analytical methods. As UN Secretary-General Ban Ki-moon remarked, "our traditional 20th century tools for tracking

international development cannot keep up." (United Nations, 2011) This paper responds to this situation by discussing the emergence of BD4D as an increasingly 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. While our discussion emphasizes a number of ways in which development problems are approached through big data analysis, a range of issues require further scrutiny. In particular, we see the *organizational arrangements* and *forms of knowledge production* related to BD4D as important paths to explore empirically. The growing reliance on big data analyses requires intricate partnerships and other inter-organizational relations that make the sourcing and aggregation of disparate forms of data possible. The development and institutionalization of such arrangements remain unexplored to date and affect the four dimensions of an analytics of government discussed in this paper. Empirical studies along these lines should focus on the creation and nature of inter-organizational relations and partnerships, negotiations over data ownership, formatting and lineage, and how professions, organizational identities and boundaries are reconfigured as a result of such developments.

Future empirical research also needs to look at the forms of knowledge production and (in)visibilities that BD4D entails. Big data analyses produce particular kinds of visibility and invisibility that reconfigure not only governance, but also "quotidian living" (Kallinikos, 2011). Empirical studies need to explore how big data intersects with other forms of knowledge production (both qualitative and quantitative) when making objects and subjects seeable, knowable and governable. 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, as they show how the current disconnect between big data analyses and the traditional statistics community can be overcome.

REFERENCES

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

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

Bughin J, Livingston J and Marwaha S (2011) Seizing the potential of "Big Data". *McKinsey Quarterly* 4: 103–109.

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 (accessed 5 January 2014).

Chan EH, Sahai V, Conrad C and Brownstein JS (2011) Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS Neglected Tropical Diseases* 5: e1206.

Chen H, Chiang RHL and Storey VC (2012) Business intelligence and analytics: from big data to big impact. *MIS Quarterly* 36: 1165-1188.

Chinn MD and Fairlie RW (2010) ICT use in the developing world: an analysis of differences in computer and internet penetration. *Review of International Economics* 18: 153-167.

Cukier K and Mayer-Schönberger V (2013) The rise of big data. *Foreign Affairs 92*(3): 27-40.

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.

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 (1984) Space, knowledge, power. In: Rabinow P (Ed.) *The Foucault reader.* London: Penguin, 239–256.

Foucault M (1991a) Governmentality. In: Burchell G, Gordon C and Murphy D (Eds.) *The Foucault effect: Studies in governmentality*. London: Harvester, 87–104.

Foucault M (1991b) Questions of method. In: Burchell G, Gordon C and Murphy D (Eds.), *The Foucault effect: Studies in governmentality*. London: Harvester, 73–86.

Frankel F and Reid R (2008) Distilling meaning from data. *Nature 455*: 30.

Gabriel Y (2008) Against the tyranny of PowerPoint: Technology-in-use and technology abuse. *Organization Studies* 29: 255–276.

Gantz J and Reinsel D (2012) *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far East.* Available at:

http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf (Accessed 28 October 2013).

Gaver WW (1991) Technology Affordances. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching Through Technology*: 79–84.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS and Brilliant L (2009) Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–4.

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

Gouldson A and Bebbington J (2007) Corporations and the governance of environmental risk. *Environment and Planning C: Government and Policy 25*(1), 4–20.

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 (accessed 15 June 2013).

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

Ilcan S and Phillips L (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.

Jaeger HM (2010) UN reform, biopolitics, and global governmentality. *International Theory* 2: 50-86.

Kallinikos, J (2011) The new everyday, *Telos,* available at http://www.telos-eu.com/en/culture-and-society/the-new-everyday.html, accessed March 28th, 2014

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

Kohut A, Wike R, Horowitz JM, Simmons K, Poushter J, Barker C, Bell J, Gross EM (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 online at http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, D, Kennedy, R. and Vespignani, A (2014) The Parable of Google Flu:

Traps in Big Data Analysis, *Science, vol 343: 1203-1205*

Liu B (2010) Sentiment analysis and subjectivity. In: Indurkhya N and Damerau FJ (Eds.) *Handbook of Natural Language Processing*. Boca-Raton, FL: Taylor & Francis, 627–665.

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 (accessed 27 August 2013).

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

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

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

Miller P and Rose N (1990) Governing economic life. *Economy and Society* 19: 1-3

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

Nelson S (2008) Big data: The Harvard computers. *Nature* 455: 36–37.

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

Norris P (2001) *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge: Cambridge University Press.

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 (accessed 17 October 2013).

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

Rose N (1999) *Powers of freedom: Reframing political thought*. Cambridge: Cambridge University Press.

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 (accessed 28 August 2013).

Sauder M and Espeland WN (2009) The discipline of rankings: Tight coupling and organizational change. *American Sociological Review* 74: 63–82.

Spence LJ and Rinaldi L (2012) Governmentality in accounting and accountability: A case study of embedding sustainability in a supply chain. *Accounting, Organizations and Society* (forthcoming).

Takeda MB and Helms MM (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(2): 204-217.

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

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 (accessed 15 October 2013).

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.

United Nations Children's Fund (UNICEF). 2013. *Tracking anti vaccination sentiment in Eastern European social media networks*. New York, NY: UNICEF.

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

World Economic Forum (2012) *Big Data, big impact: New possibilities for international development*. Geneva: World Economic Forum.

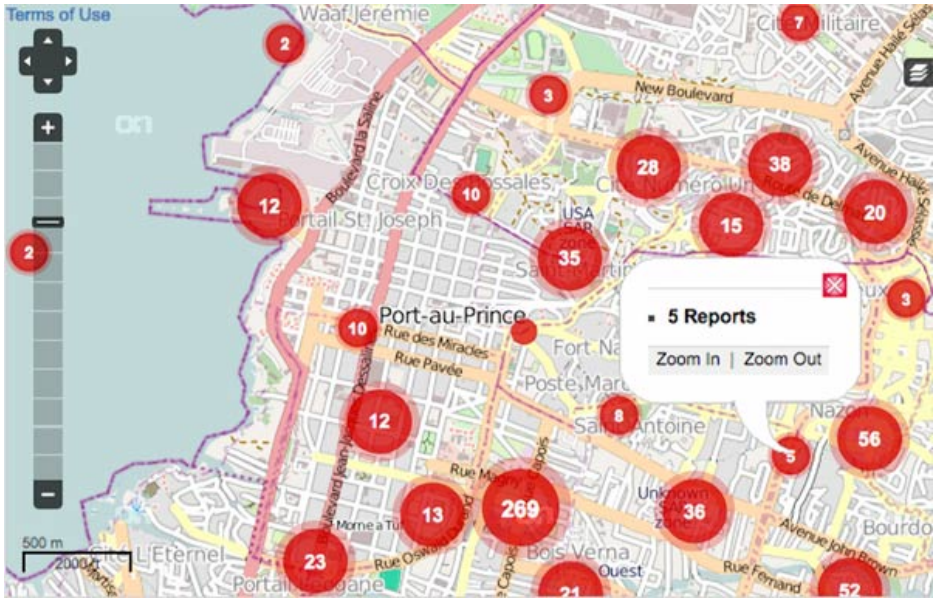
Table 1: Selected Initiatives on Big Data for Development

Organization	Year of Creation	Big Data for Development Focus	Website
United Nations (UN) Global Pulse	2009	A hub within the UN to stimulate partnerships between organizations that have access to big data and relevant UN agencies. UN Global Pulse itself has three labs for developing and testing different uses of big data within the context of development (e.g. on using tweets to predict crisis-related stress).	www.unglobalpulse.org
Ushahidi	2008	A platform facilitating the crowdsourcing of information to create real-time crisis maps. Ushahidi has been used, among other things, after the Haitian earthquake to coordinate relief efforts, in the DR Congo to track unrest, and in Kenya to map violence after the post-election fallout in 2008.	www.ushahidi.com
InSTEDD	2006	InSTEDD aims at improving the communication flow between multiple parties in the context of emergencies, diseases, and disasters. InSTEDD has developed "Resource Map" – a tool that uses mobile phone data to track medical supplies and personnel, food prices, and environmental conditions (e.g. air and water quality)	www.instedd.org
Google Flu/Dengue Trends	2007	Provides near real-time estimates of flue/dengue activity based on an analysis of aggregated search queries. The estimates are compared against historical data (i.e. baseline) to then assess the current activity level (minimal, low, moderate, high, intense). Data is available for over 25 countries for flu trends and three countries for dengue trends (as of 2013).	www.google.org/flutrends www.google.org/denguetrends
DataKind	2012	A project-based platform bringing together data scientists with organizations that have well-defined objectives for using data for development-related purposes. One project (together with the World Bank) aimed at integrating real-time data from different sources to enhance inflation calculations and to better predict local food crises.	www.datakind.org

Figure 1: Crisis Monitors Used for Exploring Topics Behind Price Spikes in Indonesia

Source: UN Global Pulse and Crimson Hexagon (2011)

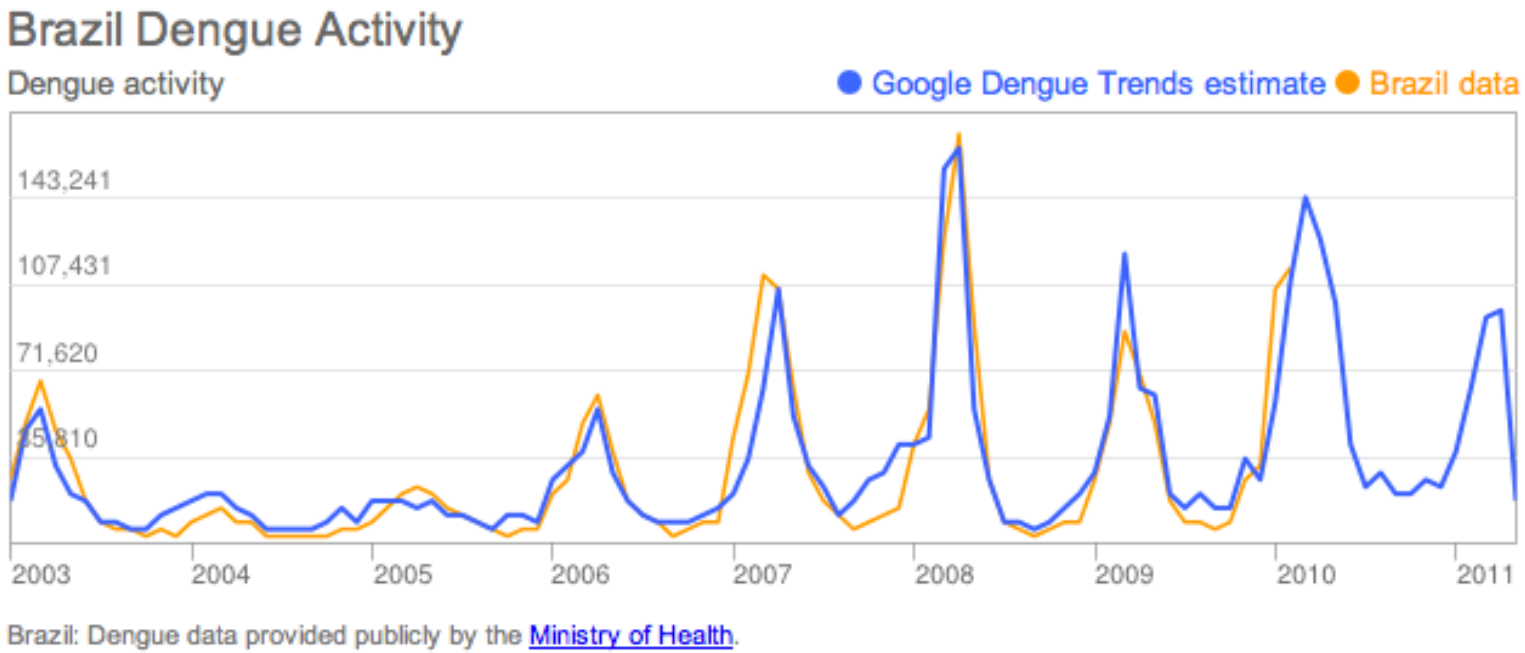
Figure 2: Ushahidi Haiti Crisis Map



Note: Numbers reflect the number of reports in a specific area. Users were allowed to zoom in further to see the details of the individual reports.

Source: Ushahidi Haiti Project

Figure 3: Google Dengue Trend Estimates Plotted Against Official Data by the Brazilian Ministry of Health



Source: Google.org (http://www.google.org/denguetrends)