K4D Knowledge, evidence and learning for development

Benefits and risks of Big Data Analytics in fragile and conflict affected states

Iffat Idris
GSDRC, University of Birmingham
17 May 2019

Question

What are the benefits and risks of big data analytics in fragile and conflict-affected states?

Contents

1. Summary

Big Data is an umbrella term for the large amounts of digital data continually generated by the global population. The main sources are data exhaust (largely from use of mobile phones), online information (e.g. social media), physical sensors (e.g. satellite imagery) and crowdsourced data (from citizens). Big Data for Development refers to sources of Big Data relevant to policy and programming of development programmes. Such data has the following features: digitally generated, passively produced, automatically collected, geographically or temporally trackable, and continuously analysed. Big Data analytics refers to the process of turning raw data into actionable information. The biggest source of Big Data is data exhaust, much of which is held by the private sector. Donor interest in Big Data for Development has increased hugely in recent years.

Big Data can support global development in three broad areas:

a) **Early warning** – Early detection of anomalies can enable faster responses to populations in times of crisis;
b) **Real-time awareness** – Fine-grained representation of reality through Big Data can inform the design and targeting of programmes and policies;
c) **Real-time feedback** – Adjustments can be made possible by real-time monitoring of the impact of policies and programmes.

Big Data is particularly useful for fragile and conflict affected states (FCAS):

- It can help overcome the challenges faced in carrying out traditional research in such contexts, due to insecurity, lack of access, population movement, etc.;
- The speed with which Big Data is generated can reduce the time lag between the start of a trend/development and when governments and other authorities are able to respond;
- Big Data allows information to be associated with individuals and their locations;
- It can also shed light on disparities in society that might otherwise be hidden, e.g. about women and girls;
- Crowdsourcing allows individuals to contribute to the information gathering process, making it more democratic and transparent;
- Big Data analytics can be a less costly and more efficient method of gathering information than traditional research approaches such as surveys.

Overall, Big Data enables development actors to strengthen decision-making processes, improve service delivery, elicit meaningful citizen participation, and increase responsiveness in humanitarian services. However, the literature stresses that Big Data is not a replacement for traditional development research and data, but rather the two should be used together.

Examples of the use of Big Data for Development include:

- Monitoring disease prevalence, predicting disease outbreaks, and the spread of epidemics through populations, largely through cell phone data records, as well as remote imaging;
- Estimating GDP from light emissions at night through remote sensing;
- Gaining insights into how people perceive issues related to food, fuel, housing the economy through analysis of social media (Twitter data);

- Predicting crop yields and thus food security through remote satellite imagery;
- Monitoring Twitter for spikes in the volume of messages about earthquakes, to enable early verification of earthquakes, and locate the epicentre;
- Using crowdsourcing through a mobile phone app to obtain citizen feedback on water quality.
- Using satellite imagery to greatly improve the spatial resolution of existing data on girls' stunting, women's literacy, and access to modern contraception in a number of developing countries.
- Analysis of tweets from more than 50 million Twitter accounts across the world to understand the differing priorities of women and men on topics related to sustainable development.

There are a number of challenges and risks associated with use of Big Data:

- **Privacy and security:** this is important for individuals, as well as companies and states. Privacy concerns have implications for all aspects of Big Data for Development: data acquisition, data storage and retention, use and presentation. Protecting privacy is particularly important in the context of FCAS as failure to do so can have serious security implications.
- **Access to Big Data:** much of Big Data is held by private sector corporations, who may be reluctant to share it for various reasons. There can also be institutional and technological challenges in sharing data.
- **Methodological challenges in analytics:** difficulties in sentiment analysis (opinion mining) and text mining; falsification of data; gap between perceptions and facts; sampling selection bias (people generating digital data might not be representative of the wider population); apophenia or seeing patterns and correlations where none actually exist; even where there is correlation, it might not signify causation.
- **Capacity challenges in analytics:** Making use of Big Data requires infrastructure (hardware and software) as well as capacity and skills. The resources required (computers, servers, etc.) are costly and the technical capacities require highly skilled labour – both are mostly found in the developed world.
- **Digital divide:** The expansion of Big Data carries with it the risk of digital divide growing between those who are able to generate and use digital data and those who are not. This can be found between countries and within countries.

Recommendations in the literature to overcome these challenges include: ensuring proper regulation and protection of data so people cannot be identified; data philanthropy by private firms to share data with others; verifying Big Data findings with local knowledge and on-ground (traditional) research; and pooling resources and capacity for Big Data analytics.

Big Data for conflict prevention has obvious relevance for fragile and conflict affected states. As well as the volume of data generated and the speed with which it is produced, it is the insights into what people think that make Big Data especially useful in preventing violent conflict. Examples of Big Data applications for conflict prevention include analysis of migration patterns using cellphone data records and remote imaging; social media analysis to identify hate speech, and issues causing stress in society; modelling and predicting social upheaval and revolution by tracking food prices; and using crowdsourced data to give real-time information about events on the ground. While Big Data is increasingly effective for *documenting* conflict situations, it is yet to

be proven in practice as effective for *predicting* conflicts. The challenges involved in using Big Data for conflict prevention are similar, albeit often more intense, than those generally involved in using Big Data for Development. However a fundamental challenge is the disconnect between obtaining information and acting on it.

This review drew on academic and grey literature. This was largely gender-blind.

2. What is Big Data for development?

Definitions

Big Data

Big Data is an umbrella term for the large amounts of digital data continually generated by the global population. Generated by a range of sources, the spread of technology means the amount of digital data available has increased on a massive scale (UN Global Pulse, 2013; GIZ, 2017). The exponential rate of increase in the quantity of digital data is likely to continue in years to come: by 2021 8.3 billion mobile phones (with over 50% smartphones) are expected to be in use, and the Internet of Things could increase from 8.4 billion objects in 2017 to 20 billion in 2020 (GIZ, 2017: 11). Indeed, a 2013 report predicted that it would increase by an annual 40% (UN Global Pulse, 2013: 1). Big Data – characterised by its high-volume, high-velocity and high-variety – is distinct from Open Data, which refers to data that is free from copyright and can be shared in the public domain (UN Global Pulse, 2013: 2). Much of Big Data is actually held by the private sector.

Big Data for Development

Big Data for Development refers to the identification of sources of Big Data relevant to policy and programming of development programmes (UN Global Pulse, 2013: 3). It thus differs from both traditional development data (e.g. survey data, official statistics) and the wider concept of Big Data. In general sources of Big Data for Development are those which can be analysed to gain insight into human well-being and development. They generally share some or all of the following features (UN Global Pulse, 2012: 15):

- **Digitally generated:** Data is created digitally (as opposed to being digitised manually), and can be manipulated by computers;
- **Passively produced:** Data is a by-product of our daily lives or interaction with digital services;
- **Automatically collected:** A system is in place that automatically extracts and stores the relevant data as it is generated;
- **Geographically or temporally trackable:** E.g. mobile phone location data, or call duration time;
- **Continuously analysed:** Information is relevant to human well-being and development, and can be analysed in real time[1].

[1] 'Real time' does not always mean immediately; rather it can be understood as information which is produced and made available in a relatively short and relevant period of time, and information which is made available within a timeframe that allows action to be taken in response, i.e. creating a feedback loop.

Big Data analytics

Big Data analytics is actually the proper term for Big Data, as it 'goes far beyond the increasing quantity and quality of data, and focuses on analysis for intelligent decision-making' (Hilbert, 2016: 6). Global Pulse (2013: 4) define Big Data analytics as 'a type of quantitative research that examines large amounts of data to uncover hidden patterns, unknown correlations and other useful information'. Tools and methodologies are needed to convert massive quantities of raw data – imperfect, complex and often unstructured – into actionable information (Global Pulse, 2013). In the words of a Stanford academic: 'data is the new oil: like oil, it must be refined before it can be used' (UN Global Pulse, 2012: 13).

UN Global Pulse (2012: 13) identify three requirements to make effective use of Big Data: a) *availability* of raw data; b) *intent* to utilise it; and c) *capacity* to understand and use data. These requirements are discussed below in the context of challenges in Big Data analytics.

Sources of Big Data

The literature identifies four categories of sources of Big Data for Development (UN Global Pulse, 2012: 16; Vaitla, 2014):

- **Data Exhaust** – passively collected transactional data from people's use of digital services like mobile phones, purchases, web searches, etc., and/or operational metrics and other real-time data collected by UN agencies, NGOs and other aid organisations to monitor their projects and programmes (e.g. stock levels, school attendance) These digital services create networked sensors of human behaviour;
- **Online Information** – web content such as news media and social media interactions (e.g. blogs, Twitter), news articles obituaries, e-commerce, job postings. This approach considers web usage and content as a sensor of human intent, sentiments, perceptions, and want;
- **Physical Sensors** – satellite or infrared imagery of changing landscapes, traffic patterns, light emissions, urban development and topographic changes, etc.. This approach focuses on remote sensing of changes in human activity;
- **Citizen Reporting or Crowd-sourced Data** – Information actively produced or submitted by citizens through mobile phone-based surveys, hotlines, user-generated maps, etc. While not passively produced, this is a key information source for verification and feedback.

Of the four categories, data exhaust accounts for the biggest share, while the major source of data exhaust in the developing world is data arising from mobile phone use (Vaitla, 2014: 2). This includes data on anonymised caller and receiver phone IDs, the start and end times of calls, call duration, the location of the caller and receiver, SMS and multimedia content, and airtime expense records (amount of purchase, time of purchase, existing balance, phone user's ID, nearest cell tower location at time of purchase) (Vaitla, 2014: 2).

Online activity data is dominated by information from Twitter feeds and Google searches. Twitter is the main source of sentiment and opinion data due to its public accessibility, the amount of information available, user diversity and the range of topics discussed (Vaitla, 2014: 6). Sentiment analysis sheds light on the underlying attitudes of social media users that give rise to the ideas and emotions expressed in feeds, how these ideas move through social networks, are opposed or confirmed, and evolve in content (Vaitla, 2014: 7).

An enormous amount of data from sensing technologies such as satellites has been collected for the past several decades, but the launch of new satellites by middle-income countries over the last few years has expanded the quantity of data even more. Furthermore, 'increases in the computing power available to analyse these massive data sets, and new methodologies developed to take advantage of this computer power, open up exciting possibilities for research' (Vaitla, 2014: 8).

Main actors involved

Vaitla (2014) highlights a big difference between traditional development research data and Big Data. The former is most commonly generated by government statistical agencies, multilateral agencies and a few specialised NGOs, and the data rests with them. Academics generally work with either the big datasets controlled by these public and non-profit entities, or with smaller datasets they generate themselves specifically for research purposes. Governments are increasingly involved in partnerships with academics, as research aims to feed more directly into the policy process. The private sector has played a relatively small role in traditional development research.

By contrast, data exhaust – the biggest source of Big Data – is often owned by the private sector, specifically by mobile phone companies. Similarly, data from online activity (notably Twitter, Google searches) is largely generated through the private sector, albeit this data, as well as remote sensing and crowdsourced information, is often publicly available. Because of the capacity requirements for making sense of the vast, complex body of raw data, 'academics currently have a high degree of influence in how Big Data is actually utilised' (Vaitla, 2014: 1). However, traditional development agencies also have a role: UN Global Pulse has been active in marketing the potential of Big Data for development, and initiating research initiatives on its own. The United Nations Children's Fund (UNICEF) has launched several innovative social media and crowdsourcing pilot projects. Donor interest in Big Data for development has increased hugely in recent years.

3. Benefits of Big Data Analytics for FCAS

Benefits and applications

The combination of the size, speed and nature of Big Data make it highly valuable to affect development outcomes (UN Global Pulse, 2012: 38). According to Global Pulse (2013: 7), 'the question is no longer if Big Data can provide insights useful to global development and resilience, but how'. Global Pulse (2012; 2013) identify three broad areas in which Big Data can support global development:

d) **Early warning** – Early detection of anomalies can enable faster responses to populations in times of crisis;
e) **Real-time awareness** – Fine-grained representation of reality through Big Data can inform the design and targeting of programmes and policies;
f) **Real-time feedback** – Adjustments can be made possible by real-time monitoring of the impact of policies and programmes.

A World Economic Forum (WEF, 2012: 4) paper gives a similar categorisation of uses of Big Data for Development:

- Faster outbreak tracking and response;

- Improved understanding of crisis behaviour change;
- Accurate mapping of service needs;
- Ability to predict demand and supply changes.

Big Data for Development has potential applications in numerous sectors, notably health, education, financial services and agriculture – examples are discussed below. It can be especially useful in fragile and conflict affected states where, for multiple reasons (insecurity, lack of capacity, population movements, etc.), availability of traditional data (e.g. official statistics) tends to be very limited. The speed with which Big Data is generated can reduce the time lag between the start of a trend/development and when governments and other authorities are able to respond; it can also reduce the knowledge gap about how people respond to these trends (WEF, 2012: 4).

> Real-time data has the potential to give development actors the means to uncover anomalies, respond to issues as they arise, improve internal coordination, optimise resource allocation, react to citizen feedback, and anticipate trends and future events (GIZ, 2017: 11).

Big Data allows information to be associated with individuals and their locations (Vaitla, 2014). It can also shed light on disparities in society that might otherwise be hidden, e.g. about women and girls who often work in the informal sector at home and have limited mobility (UN, nd). The use of Big Data in a range of projects highlights its potential to close the global gender data gap (Data2x, 2017). Crowdsourcing allows individuals to contribute to the information gathering process, making it more democratic and transparent (WEF, 2012).

Big Data analytics can be a less costly and more efficient method of developing market intelligence for large organizations like the World Bank: 'The Bank already spends millions of dollars each year on statistical analysis of the needs of the poor. Smarter data collection and analysis could free resources for use in economic development efforts' (WEF, 2012: 5). This is particularly important in times of constrained government resources and reduced foreign aid.

In sum (GIZ, 2017: 14):

> Digital data offers manifold opportunities to development by enabling development actors to strengthen decision-making processes, improve service delivery, elicit meaningful citizen participation, and increase responsiveness in humanitarian services, among others.

The literature stresses that Big Data is not a replacement for traditional development research and data (Escobal et al, 2018; UN Global Pulse, 2012). For example, Vaitla (2014: 12) explains that, far from replacing traditional data systems, 'for the foreseeable future the utility of big data will depend on the creative combining of big data and traditional data sets to analyse development phenomena'. Conventional survey methods are often needed to validate the representativeness of Big Data – discussed below – or to identify the nature and magnitude of biases within Big Data sets (Vaitla, 2014: 12).

Examples of Big Data for Development: public health

Global Viral Forecasting Initiative – Analysis of mobile phone and internet data could lead to huge gains in public health. The Global Viral Forecasting Initiative (GVFI), based in San Francisco, uses advanced data analysis on information mined from the internet to identify comprehensively the locations, sources and drivers of local disease outbreaks before they

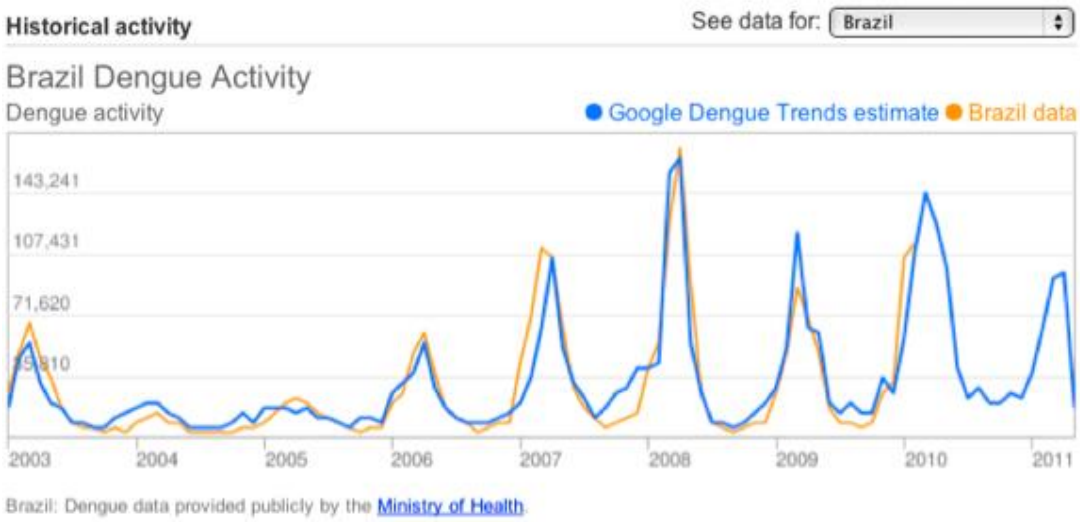
become global epidemics (WEF, 2012: 5). GVFI claim the technique can successfully predict outbreaks up to a week ahead of global bodies such as the World Health Organisation which rely on traditional techniques and indicators (WEF, 2012: 5).

Migration and disease spread in Kenya and Haiti – Geographic mobile phone records from rural Kenya have been used to provide detailed travel and migration patterns in low-income settings to understand the spread of malaria and infectious diseases (Hilbert, 2016: 11). Following the 2010 Haiti earthquake researchers at the Karolinska Institute and Columbia University showed that mobile data patterns could be used to understand the movement of refugees and the consequent health risks posed by these movements. Researchers obtained data on the outflow of people from Port-au-Prince after the earthquake by tracking the movement of nearly two million SIM cards in the country (Vaitla, 2014: 3). They were able to accurately analyse the destination of over 600,000 people displaced from Port-au-Prince, and they made this information available to government and humanitarian organisations dealing with the crisis (WEF, 2012: 5). Later that year, a cholera outbreak struck the country and the same team used mobile data to track the movement of people from affected zones. Aid organisations used this data to prepare for new outbreaks. 'Mobile phones give researchers the ability to quantify human movement on a scale that wasn't possible before' (UN Global Pulse, 2012: 20). A retrospective analysis of the 2010 cholera outbreak in Haiti conducted by researchers at Harvard and MIT demonstrated that mining Twitter and online news reports could have provided health officials a highly accurate indication of the actual spread of the disease with two weeks lead time (UN Global Pulse, 2012: 37; Vaitla, 2014: 4).

Disease prevalence – Remote sensing systems can be used to predict disease prevalence. They can be used to predict the degree of both inter-year and intra-year risk of illness, and thereby help mobilize preventative and curative resources well in advance of actual morbidity and mortality (Vaitla, 2014: 8). Malaria has been by far the most studied disease through remote sensing – sensing research on malaria vector abundance dates back over two decades – but a wide variety of other illnesses have also been analysed, including lyme disease, cholera, meningitis, dengue, Rift Valley fever, schistosomiasis, West Nile fever, and even obesity (Vaitla, 2014: 8). In the last decade or so, thanks in part to increased computing power, this research has been used in the policy process through the creation of epidemic early warning systems and other initiatives (Vaitla, 2014: 8).

Google Dengue Trends – Google Dengue Trends uses aggregated Google search data to estimate current dengue activity around the world in near real-time (UN Global Pulse, 2012: 21). Google found a close relationship between how many people search for dengue-related topics and how many people actually have dengue symptoms – Figure 1 shows this for Brazil. Google compared their query counts with traditional dengue surveillance systems and found that many search queries tended to be popular exactly when dengue season was happening. By counting how often those search queries are seen, Google can estimate how much dengue is circulating in different countries and regions around the world.

Figure 1: Correlation between Google dengue search data and Ministry of Health dengue surveillance in Brazil



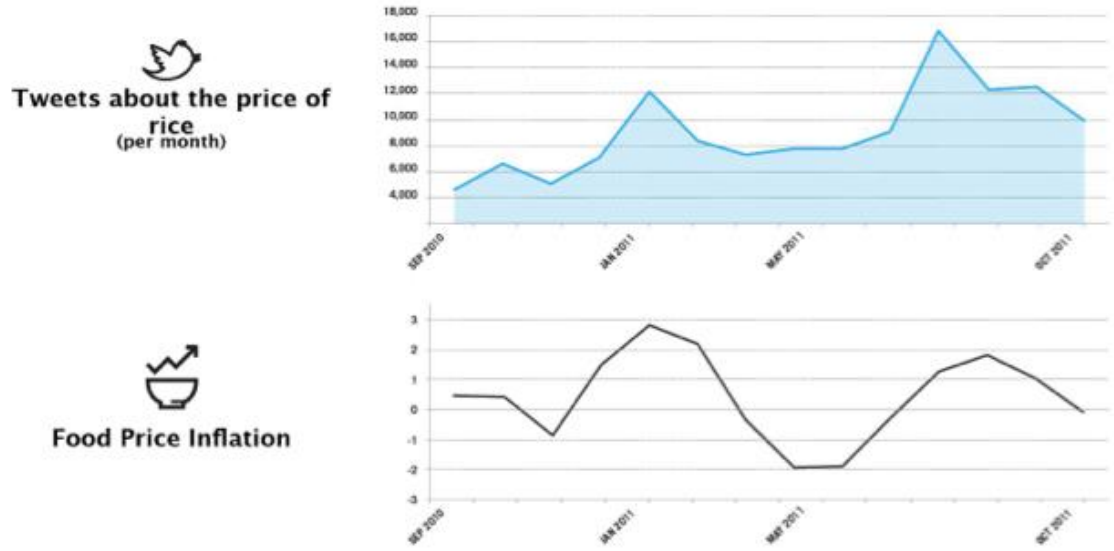
Reproduced with kind permission of: UN Global Pulse, 2012, p.21

Examples of Big Data for Development: socioeconomic well-being

GDP estimates – A country's GDP can be estimated in real time by measuring light emissions at night through remote sensing (UN Global Pulse, 2012: 20).

Food security – To evaluate the effectiveness of harnessing Big Data for Development, UN Global Pulse has worked on several research projects in collaboration with public and private partners. In one, the agency partnered with social media analysis firm Crimson Hexagon to see if Twitter data could be used to provide insight about how people in Indonesia perceive issues related to food, fuel, housing and the economy in Indonesia. Over 14 million tweets related to food, fuel, and housing were analysed. The researchers found that the number of tweets discussing the price of rice in Indonesia closely matched the official inflation statistics (see Figure 2).

Figure 2: Tweets about the price of rice vs. actual price of rice in Indonesia



Reproduced with kind permission of UN Global Pulse, 2012, p.38

In another project, UN Global Pulse partnered with the UN Millennium Campaign and DataSift to scan Twitter for the most commonly discussed development-related topics worldwide. The results suggested that, in contrast to surveys that ask about long-term priorities, analysis of Twitter feeds enables a unique understanding of 'daily hopes and grievances' (Vaitla, 2014: 6).

Remote sensing in food security forecasting and development mapping – Remote sensing is being used for food security forecasting, based on satellite sensing of vegetation density. The data gathered by satellites enables crop yields to be predicted well before the actual harvest (Vaitla, 2014: 8). When combined with household-level information on factors like family size, available labour, input and output costs, and so on, this data can help evaluate the risk of impending food insecurity. The main use of remote sensing in development is the mapping of environmental phenomena and human infrastructure, including vegetation, water bodies, transport networks, and land use patterns (Vaitla, 2014: 8).

Examples of Big Data for Development: real-time awareness

Early warning about earthquakes - The United States Geological Survey (USGS) has developed a system that monitors Twitter for significant spikes in the volume of messages about earthquakes (UN Global Pulse, 2012: 37). Location information is then extracted and passed on to USGS's team of seismologists to verify that an earthquake occurred, locate its epicentre and quantify its magnitude. 90% of the reports that trigger an alert have turned out to be valid (UN Global Pulse, 2012: 37).

Crowdsourcing to identify earthquake damage - Following the 2010 Haiti earthquake, the NGO Ushahidi set up a centralised text messaging system to allow cell-phone users to report on people trapped under damaged buildings. This information helped the US military and other emergency responders to find individuals in need; indeed, in the early days following the disaster Ushahidi was the only source of aggregate geospatial information, processing at least 40,000

reports and mapping nearly 4,000 individual events (Vaitla, 2014: 10). Analysis of the crowdsourcing data found that the concentration of aggregated text messages was highly correlated with areas where the damaged buildings were concentrated (UN Global Pulse, 2012: 23). The NGO claimed the results were evidence of the system's ability to 'predict, with surprisingly high accuracy and statistical significance, the location and extent of structural damage post-earthquake' (UN Global Pulse, 2012: 23).

Crowdsourcing to obtain citizen feedback on water quality - Vaitla (2014: 11) notes that crowdsourcing is gaining favour among donors. USAID provided funding for a non-profit organization, mWater, to develop a mobile phone application in Tanzania that helps citizens perform water quality tests and upload this information to a database that maps water sources (Vaitla, 2014: 11).

Examples of Big Data for Development: closing the gender data gap

Geospatial data to gather social and health data on women and girls – Satellite imagery was used to greatly improve the spatial resolution of existing data on girls' stunting, women's literacy, and access to modern contraception in Bangladesh, Haiti, Kenya, Nigeria and Tanzania (Data2X, 2017: 1). The approach takes advantage of the fact that many types of health and social data – such as child stunting, literacy and access to contraception – are correlated with geospatial phenomena that can be mapped in great detail across entire countries using satellite imagery. The study cited generated a series of highly detailed maps that clearly showed landscapes of gender inequality (Data2X, 2017).

Use of data exhaust to create economic portraits of women – A project in Latin America used anonymised credit card and cell phone data to describe patterns of women's expenditure and mobility in a major metropolis. The information was used to create portraits of economic lifestyles – patterns of behaviour that illustrate the needs and priorities of women. Over a longer timeframe, such data could also reveal signals about how women cope with a wide range of environmental and economic shocks and stressors.

Analysis of tweets to quantify women's concerns on development issues – A project implemented globally developed a tool for automatically identifying the sex of Twitter users (from users' names and, if needed, pictures from Twitter profiles). The tool was tested on more than 50 million Twitter accounts across the world to understand the differing priorities of women and men on topics related to sustainable development (Data2X, 2017).

4. Challenges and risks

Privacy (and security)

Privacy is defined as 'the right of individuals to control or influence what information related to them may be disclosed' (UN Global Pulse, 2012: 24). In the context of Big Data it can also be understood in the broader sense: encompassing companies' wish to protect their competitiveness, consumers and reputation, and states' wish to protect their sovereignty and citizens. Privacy is a fundamental human right: 'without privacy, safety, diversity, pluralism, innovation, our basic freedoms are at risk' (UN Global Pulse, 2012: 24). In the context of fragile

and conflict affected states, the failure to protect privacy can have serious security implications for people.

Privacy concerns have implications for all aspects of Big Data for development: data acquisition (are people aware of the data they are generating, and what it will be used for?), storage and retention, use and presentation (UN Global Pulse, 2013: 6). Because Big Data is the product of unique patterns of behaviour of individuals, removal of explicit personal information may not fully protect privacy – combining multiple datasets could lead to the re-identification of individuals or groups of individuals, exposing them to potential harm (UN, nd).

To address these risks it is important to have suitable legal frameworks, ethical guidelines and technological solutions for protected data sharing.

Access

As seen, much of Big Data is held by private sector corporations. They may be reluctant to share data because of concerns about competitiveness, their customers' privacy, their own reputation and liability, a culture of secrecy and/or absence of the right incentive and information structures (UN Global Pulse, 2012: 25). There can also be institutional and technical challenges in sharing data, e.g. when data is stored in places and ways that make it difficult to be accessed and transferred, issues with inter-comparability of data and inter-operability of systems.

UN Global Pulse has put forth the concept of 'data philanthropy' to encourage private sector sharing of data. Through data philanthropy 'corporations take the initiative to anonymize (strip out all personal information) their data sets and provide this data to social innovators to mine the data for insights, patterns and trends in real-time or near real-time' (UN Global Pulse, 2012: 25). Another arrangement that is emerging is 'data collaboratives' whereby those involved share data assets and combine their expertise and/or tools to solve specific public problems (GIZ, 2017: 12).

Analytics

There are a number of risks or challenges involved in the analytics of Big Data – the process of extracting relevant information – which relate to one, infrastructure and capacity to conduct analysis (see 'Digital Divide' below) and, two, methods of analysis (UN Global Pulse, 2013: 7):

- **Sentiment analysis (or opinion mining) -** Refers to the study of emotions and opinions expressed in digital messages and translating those sentiments to hard data. Quantifying moods and intents is difficult, and obstacles such as slang, sarcasm, hyperbole, and irony may impede data analysis.
- **Text mining –** This goes beyond sentiment analysis to the extraction of keywords and events. The difficulty here is assessing the true significance of the statements in which the facts are reported.
- **Falsification -** Data can be false or fabricated with the intention of providing misleading information.
- **Perceptions versus facts -** Perceptions are not necessarily accurate and may differ even significantly from actual facts. For instance, this happened with Google Flu Trends, an analytics platform that was meant to predict actual flu, but instead proved useful only for general public health surveillance. Without recognising this key insight, doctors using Google Flu Trends may be inclined to overstock vaccines or misdiagnose their patients.

- **Sampling selection bias -** The people who use mobile or digital services may not be a representative sample of the larger population considered. 'Depending on the type of data, one expect younger or older, wealthier or poorer, more males than females, and educated or uneducated individuals to account for a disproportionate share of (data) producers' (UN Global Pulse, 2012: 29).
- **Apophenia -** Seeing patterns and correlations where none actually exists is a risk, and the massive amount of data available for analysis intensifies the search for interesting correlations that may not actually exist.
- **Correlation does not mean causation -** Even where an actual correlation is found, it does not necessarily signify that there is a causation link between the data and theory and context are relevant to reduce the risk of self-fulfilling prophecies.

The methodological challenges listed above do not undermine the utility of Big Data for Development; rather, they point to the need for external validation of findings through traditional research methods and through local knowledge and understanding of the context. 'The promise of Big Data for Development is, and will be, best fulfilled when its limitations, biases and ultimately features, are adequately understood and taken into account when interpreting the data' (UN Global Pulse, 2012: 36).

Digital divide

The expansion of Big Data carries with it the risk of digital divide growing between those who are able to generate and use digital data and those who are not.

Even though mobile phone and smartphone use and internet access is increasing globally, the availability and types of digital data can vary significantly from country to country. Countries with high mobile phone and internet penetration rates produce more data directly generated by citizens, while those with large aid communities produce more programme-related data (UN Global Pulse, 2013: 6). Similar gaps also exist within countries, for example between rural and urban communities, across age groups and economic brackets, and by gender. Escobal et al (2018: 13) note that asymmetry in access to information technologies disproportionately affects less-advantaged groups, hampering their representation in the data.

> Major gaps are already opening up between the data haves and have-nots….Many people are excluded from the new world of data and information by language, poverty, lack of education, lack of technology infrastructure, remoteness or prejudice and discrimination (UN, nd).

Making use of Big Data requires infrastructure (hardware and software) as well as capacity and skills. The resources required (computers, servers, etc.) are costly and the technical capacities require highly skilled labour that is often at a premium in the private sector (Escobal et al, 2018: 13). Unequal distribution of these can lead to a digital – and thus development – divide (Hilbert, 2016: 8). The vast majority of Big Data hardware capacity resides in highly developed countries (Hilbert, 2016: 18). Similarly inequalities are seen in ICT spending – and software and computer service spending as a percentage of total ICT spending – and the share of software and computer service employees out of total employment: in both, the figures are higher for developed countries (Hilbert, 2016: 20-21). 'This inevitably creates a new dimension *(access to data being another)* of the digital divide: a divide in the capacity to place the analytic treatment of data at the forefront of informed decision-making and therefore a divide in data-based knowledge' (Hilbert, 2016: 32).

> Existing information inequalities are likely to be exacerbated in the coming years as some actors are in a better position that others to harness the positive developments arising from data….. those with the capacity to handle large volumes of data will be in an advantageous position, while others will lose out (GIZ, 2017: 12).

One possible approach to overcome capacity challenges is the concept of 'analytics philanthropy' whereby actors (local and international) are linked in institutional arrangements such as fellowships, technical assistance, working groups to more fully integrated structures. One model could be regional Big Data hubs, which pool together resources and personnel from various surrounding countries and institutions (Letouze et al, 2013: 26).

5. Big Data for conflict prevention

Potential uses and examples

Big Data for conflict prevention is a subset of Big Data for Development which has obvious relevance for fragile and conflict affected states. It refers to the potential use of Big Data to support conflict-prevention efforts including early warning, crisis management, conflict resolution, peacemaking, peacekeeping and all activities aimed at strengthening these (Letouze et al, 2013: 5-6). Of the major trends driving Big Data – the vastly increased volume of data, expanded ability to collect and crunch data, and the nature of the data itself – it is perhaps the latter, which holds the most promise for conflict prevention and peace building. 'For the first time, digital media – user-generated content and online social networks in particular – tell us not just what is going on, but also what people think about the things that are going on' (Himelfarb, 2014). This presents the possibility that datasets can be tapped to understand, and pre-empt, the human sentiment that underlies violent conflict.

According to one academic, the most important aspect in using data in peace and conflict studies is knowing what questions to ask; secondly, one needs the necessary data to answer those questions (Hackl, 2017). Effective use of Big Data for conflict prevention also requires cross-discipline collaboration between data experts and conflict experts who have intimate knowledge of the social, political and geographical terrain of different locations (Himelfarb, 2014).

Some of the ways in which Big Data can be used for conflict prevention, and examples of this use, are as follows (Letouze et al, 2013: 13-14; Escobal et al, 2018):

- **To track migration** – analysis of migration patterns using cellphone data records (CDRs), mapping users' unique Internet Protocol (IP) addresses, or remote imaging from satellites;
- **To identify stress and hate speech** - studying causes and expressions of concern and stress in a given community—as has been done using Twitter in Indonesia — in order to better understand and address them before they fuel grievances. Social media analysis and other tools can be used to identify when hate speech is occurring and where influencers are located;
- **To model and predict social upheavals and revolutions** - For example, mathematicians and computer scientists have developed a model that tracks food prices and makes predictions about the risk that a riot might break out. In the case of the Arab Spring the researchers claim they submitted their analysis to the US government,

warning of risks of social unrest, four days before a Tunisian fruit vendor set himself on fire – the event seen as triggering the Arab Spring

- **To give real-time information about events on the ground** - Crowdsourced data systems mobilise communities or selected individuals within communities to feed data from the ground, enabling a real-time information stream about events. Such systems were used, for example, in Kenya in anticipation of violence in the 2010 and 2013 elections, and in Kyrgyzstan during elections, when trained monitors at polling stations reported adverse events through a mobile messaging system
- **To resolve disputes over land use** - In Indonesia incomplete data about land type, usage and ownership has led to frequent conflicts between businesses and local communities over territorial rights. Citizen-generated data via drones is being used to create cartographic material that can be used to resolve disputes over land use (GIZ, 2017: 15).
- **To prevent misinformation escalating to conflict** - In Kenya an initiative uses citizen-generated data via a platform called 'Una Hakiki?' (Swahili for 'Are you sure?') to prevent conflicts. People can report potentially harmful rumours on the platform, which are then quickly validated to mitigate the escalation of ethnic conflicts based on misinformation (GIZ, 2017: 15).

The potential of Big Data for national security is reflected in the fact that the Obama administration allocated approximately US$ 250 million for Big Data projects at the Department of Defense, including US$ 60 million for research (Letouze et al, 2013: 13). However, Letouze et al (2013: 4) stress that, 'as a field of practice in the making…Big Data for conflict prevention is best characterized by its potential rather than by its track record'. They note that, 'while we are increasingly able to document what is happening (descriptive use), we remain, in the case of conflict, largely blind as to what will happen next (predictive use)' (Letouze et al, 2013: 17).

Challenges and risks

The challenges and risks associated with Big Data for conflict prevention are similar to those related to Big Data for Development, albeit often more intense. Privacy concerns, for example, are paramount in conflict situations, given the 'peculiar security risks that individuals may face in some highly dangerous environments' (Letouze et al, 2013: 20). 'If not properly regulated or protected, sensitive data in conflict zones can be used by malignant actors to target vulnerable populations' (Escobal et al, 2018: 6).

Analytical challenges can also be greater. There is also a risk that data and quantitative methods could just serve as a way to solidify preconceived arguments (Hackl, 2017). Given the typically limited access in conflict situations and risks of carrying out traditional on-the-spot evaluation, this could lead to over-reliance on remote assessment tools – which might not give a representative picture (Letouze et al, 2013). Non-representativeness of data can be especially problematic in conflict zones (Letouze et al, 2013: 22):

> If unequal access to technology – and thus to most data-generating devices – may mirror conflict fault lines (e.g. social or ethnic classes) or if it results from deliberate and targeted attempts at skewing the data….The potential consequence is that conflict prevention actors relying on these data could appear to be prejudiced against or partial to specific interest groups.

Letouze et al (2013: 18) highlight that ‘information does not equal response’, i.e. even with a better understanding of violent conflict, this might not lead to decisions or actions to tackle it because of poor institutional design and/or functioning, or lack of political will. They see this as the fundamental challenge. ‘A key dimension is bridging the decision gap. Until and unless we are ready and willing to do so, Big Data, no matter how big, will not affect outcomes and save lives’ (Letouze et al, 2013: 27).

6. References

Cederman & Weidmann (2017). 'Predicting armed conflict: Time to adjust our expectation?' *Science,* 355: 474-476 (3 February 2017). https://science.sciencemag.org/content/sci/355/6324/474.full.pdf

Data2X (2017). *Big Data and The Well-Being of Women and Girls: Applications on the Social Sciences Frontier*. https://www.data2x.org/wp-content/uploads/2017/03/Big-Data-and-the-Well-Being-of-Women-and-Girls.pdf

Escobal, L. et al (2018). *Big Data for Peace and Security*. UN Peacebuilding Support Office.

GIZ (2017). *Data for Development: What's Next? Concepts, trends and recommendations for German development cooperation*. http://webfoundation.org/docs/2018/01/Final_Data-for-development_Whats-next_Studie_EN.pdf

Hackl, A. (2017). 'Peace and Conflict Series: Can Data Bring Peace? The Gains and Caveats of Data Science in Peace and Conflict Studies'. *Global Justice Blog,* University of Edinburgh, 26 January 2017. https://www.globaljusticeblog.ed.ac.uk/2017/01/26/peace-conflict-series-2/

Hilbert, M. (2016). 'Big Data for Development: A Review of Promises and Challenges. *Development Policy Review,* 34 (1). https://www.researchgate.net/publication/286907720_Big_Data_for_Development_A_Review_of_Promises_and_Challenges

Himelfarb, S. (2014). 'Can Big Data Stop Wars Before They Happen?' *Foreign Policy*, 25 April 2014. https://foreignpolicy.com/2014/04/25/can-big-data-stop-wars-before-they-happen/

Letouze, E. (2012). 'Big Data for Development: What may determine success or failure?' OECD. https://www.oecd.org/sti/ieconomy/Session_5_Letouz%C3%A9.pdf

Letouze, E. et al (2013). 'Big Data for Conflict Prevention: New Oil and Old Fires' in Mancini, F. (ed) (2013), *New Technology and the Prevention of Violence and Conflict*. https://acuns.org/wp-content/uploads/2013/05/New-Technology-and-the-Prevention-of-Violence-and-Conflict.pdf

UN (nd). 'Big Data for Sustainable Development'. https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html

UN Global Pulse (2012). *Big Data for Development: Challenges and Opportunities.* http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

UN Global Pulse (2013). *Big Data for Development: A Primer*. http://www.unglobalpulse.org/sites/default/files/Primer%202013_FINAL%20FOR%20PRINT.pdf

Vaitla, B. (2014). *The Landscape of Big Data for Development: Key Actors and Major Research Themes*. Data2X. https://www.data2x.org/wp-content/uploads/2017/11/LandscapeOfBigDataForDevelopment_10_28.pdf

WEF (2012). *Big Data, Big Impact: New Possibilities for International Development.* World Economic Forum. http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf

Key websites

- UN Global Pulse: www.unglobalpulse.org
- Big Data for Development (BD4D) Network: http://bd4d.net/

Suggested citation

Idris, I. (2019). *Benefits and risks of Big Data Analytics in Fragile and Conflict Affected States.* K4D Helpdesk Report 605. Brighton, UK: Institute of Development Studies.

About this report

This report is based on six days of desk-based research. The K4D research helpdesk provides rapid syntheses of a selection of recent relevant literature and international expert thinking in response to specific questions relating to international development. For any enquiries, contact helpdesk@k4d.info.

K4D services are provided by a consortium of leading organisations working in international development, led by the Institute of Development Studies (IDS), with Education Development Trust, Itad, University of Leeds Nuffield Centre for International Health and Development, Liverpool School of Tropical Medicine (LSTM), University of Birmingham International Development Department (IDD) and the University of Manchester Humanitarian and Conflict Response Institute (HCRI).