ABSTRACT

Title of Thesis: The Potential Impact of Big Data in International Development and Humanitarian Aid

Emily Jablonski, 2017

Thesis directed by: Dr. Anthony Marcum, Dr. Kentaro Toyama

This thesis attempts to answer two main questions: what effort has been made in the Big Data for development field, and what is its potential impact? An impactful project, for the extent of this paper, is defined as a project that through its use of data created a solution to a development problem that objectively improved upon the traditional method. The Big Data projects are analyzed within three categories specific to the temporal nature of the problem they are focused on: ongoing issues, intermittent issues, and sudden crises. While data-related challenges must be acknowledged by practitioners for projects to be effective, the findings show that in almost every case, Big Data solutions improved upon traditional methods and overcame the associated obstacles and criticisms. Following this analysis of Big Data's potential for impact in the development field within distinct categories, further barriers to entry are discussed and future recommendations are offered.

The Potential Impact of Big Data in International Development and Humanitarian Aid

By

Emily Jablonski

Thesis submitted to the Faculty of the College of Literature, Science, & Arts
at the University of Michigan in partial fulfillment
for the requirements for the degree
of Bachelor of Science
International Studies with Honors
2017

Thesis Committee:

Dr. Kentaro Toyama
Dr. Anthony Marcum

ACKNOWLEDGEMENTS

I would like to express my gratitude to a number of people who helped make this project possible for me. To Dr. Kentaro Toyama, thank you for your consistent support over the past academic year answering my every question, helping me form the idea for this thesis into something tangible, and for inspiration to move this idea optimistically forward into the future. To Dr. Anthony Marcum, thank you for pushing each member of the International Studies Honors Cohort to challenge each other and support each other, and for guiding us through this process. In part I owe my successful completion of this thesis to your encouragement, flexibility, and understanding. To my parents, thank you for your constant support throughout the ups and downs of my academic career. I know I would not have the opportunity to celebrate this high without you. To my closest friend, Jarrid, through all the stresses of my final undergraduate year, thank you for teaching me that I have something valuable to say.

TABLE OF CONTENTS

Chapter 1 | **Introduction**

Introduction to Big Data

Our world is becoming increasingly connected and the technologies that support this connectivity more widespread. These proliferating technologies produce and store an incomprehensible amount of data every day in each corner of the globe. Whether it takes the form of a satellite in space capturing images of weather patterns, a sea of Twitter users causing a particular hashtag to trend, an earthquake survivor in Haiti using their mobile phone, or you, feeling a little under the weather and deciding to query Google about your flu-like symptoms, data is being collected all the time through a number of different mediums. Information on this grand scale is often called Big Data.

While its definition took a number of years to gain consensus, Big Data is information typically characterized by the "Five Vs."[1] The first, volume, refers to the amount of the data being considered. Second is velocity, referring both to the speed at which the data is produced and how quickly it can be processed. Variety is third, which concerns the complexity of the data in terms of the scope of the information provided as well as how it is formatted. Fourth is value, or the

[1] Demchenko, Yuri, et al. "Addressing big data issues in scientific data infrastructure." *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013.

potential benefit the data may have for its intended use. Last is veracity, or the level of trust in the certainty of the data. Using this framework, Big Data is defined as information that embodies each of these categories at the high end of the spectrum, meaning it is sizable, fast, complex, potentially useful, and likely valid.

Such stores of data, which would have been more difficult to acquire or in some cases could not have feasibly existed at all prior to the information age, are increasingly being used by governments, corporations, and researchers to accomplish a variety of goals. Sophisticated algorithms and statistical methods have been developed and deployed on Big Data by these actors in a diverse range of fields in an attempt to create novel insights. This process, which encompasses terms like data analysis, data mining, and machine learning, is often what people mean when they refer to Big Data.

What exactly are these novel insights? For many applications of Big Data, the practice is about finding new patterns that inform the practitioner of something useful about the past, present, or future. For example, a business might feed a large amount of internal data into an analysis project to find out that employees are most likely to quit on Wednesdays, their product sells better on overcast days, and that customers in Seattle respond to ads differently than those in Boston. Other methods could have discovered these patterns, but uncovering such details by hand is time consuming. Resulting findings can be used to help assess larger current trends or possible future ones. In these few examples, the insights might help the business save time, make money, or boost the satisfaction of their workforce. Essentially, the results enable them to make smarter and often faster decisions about how to sell their product and maintain their company.

Though the above example comes from the consumer world, the applications can be generalized to many different kinds of decisions. A sports team might use real-time analytics based on past performance data to help them decide what play to run, a hospital might use data from blood tests to predict what illness a patient has and decide how to treat them, or a government might consolidate social media data to predict and prevent the actions of potential criminals based on their Facebook statuses. While its growth is relatively new, where Big Data is available, analytics can likely be applied to produce insights.

Introduction to Big Data for Development

Big Data has generated both much excitement and critique about its application to the field of international development and humanitarian aid. In their attempts to reduce human suffering and improve quality of life around the world, those at the center of international development and humanitarian aid projects sometimes turn to technologies, like those utilizing Big Data, to help accomplish their goals. As the aims behind such projects are broad, so too are the data sources they utilize. Poverty reduction programs might benefit from using microfinance data via mobile phones, drought prediction systems might find satellite imagery useful, and gender equality programs might measure women's participation in various societal activities through social media interactions. These represent just a few possibilities as examples, as Big Data is by its nature wide in scope.

Though the history of Big Data is relatively short, there does exists a small base of academic literature surrounding its use within international development and humanitarian aid projects. While there are plenty of sources concerning the advantages and disadvantages of using Big Data to help solve social problems, only a small subset of this literature consists of researchers taking part in the direct application of Big Data to help solve development problems. The literature represents a wide spectrum of beliefs about the topic, from supportive, to apprehensive, to critical. Criticisms of Big Data's use in this particular field tend to fall under two categories. First are critiques that are not necessarily specific to development, but should be considerations for any Big Data project. Objections that are more specific to issues within the development sphere encompass the second type of critique. Both are consistently noted throughout the literature.

General Big Data Criticisms

Interpretation

A central problem for any data scientist is one that math and science students learn in classrooms from a young age: correlation is not equal to causation. In short, just because valid data analytics show that two items tend to respond to each other does not mean that one is actually causing the other to change. Many purposefully absurd demonstrations of this fallacy have been undertaken to help create further understanding of this issue. One such project, which uses data to present “spurious” correlations in graph form, shows that annual US crude oil imports from Norway follow a strikingly similar path to the number of car drivers killed in collisions with trains.[2] If one wants to start a campaign to reduce train crash deaths, the main action point is likely not going to be a reduction in the importation of oil from Norway. The world is a complex place and social phenomena are particularly tricky to measure. The relation of two factors over time cannot imply a direct relationship in practice. This issue can be easier to spot outside of context than it is in practice. Data for development projects, though they

[2] Vigen, Tyler. "15 Insane Things That Correlate With Each Other." *Spurious Correlations*. N.p., n.d. Web. <http://www.tylervigen.com/spurious-correlations>.

typically involve seasoned statisticians, are not exempt from this issue, especially as they take place in a social sphere which adds an untold number of hidden variables and unknowns.[3] Big Data practitioners must be careful to not generalize results too broadly, and most all research teams in the development sphere take care to avoid this pitfall.

Hubris

One of the most consistent issues authors in the literature point to when criticizing Big Data and its applications is that some practitioners put too much faith in Big Data as a method. A sentiment exists within the community that producers of Big Data analytics often view themselves as creators of "a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."[4] Similarly, some authors express discontent that Big Data is sometimes offered as a panacea: a savior and solution to all of science's toughest problems, and a replacement for the soon-to-be obsolete traditional methods.[5]

---

[3] Letouze, Emmanuel. "Big Data for Development." UN Global Pulse, n.d. Web. <http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf>

[4] boyd, danah, and Kate Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, communication & society* 15.5 (2012): 662-679.

[5] Ebach, Malte C., et al. "Big data and the historical sciences: A critique." *Geoforum* 71 (2016): 1-4.

While this perception was significant in the literature on general Big Data applications, I did not find a theme throughout the primary research papers on Big Data for development suggesting that the authors were patronizing traditional analysis and hyper-glorifying Big Data. The authors tended to be optimistic, but did consistently note the shortcomings of their approach and the specificities of its application. This over-confident optimism seems to appear most in the marketing of business products which offer smart analytics, which is understandable from an advertising perspective. If you can convince a customer the insights you can offer them will radically change their business in a way no other method could ever compete with, they will likely want to pay you for that product, though it may be harming the perception of what Big Data can really do. Projects doing pioneering primary research in Big Data for development typically are not trying to meet a bottom line, and though noticeable elsewhere, this indicated hubris was not easy to find in this specific field.

Privacy

Another common issue deeply tied to Big Data is privacy. While there are certainly many types of Big Data, often the data concerns people as people produce much data simply by living their daily lives. For example, a morning routine might involve checking Twitter after waking up, commenting on a news story on the bus ride to work, and buying a croissant at a chain cafe for breakfast. Global data systems take note of these and many more kinds of actions people take. By simply participating in society it is difficult to avoid also participating in Big Data systems.

When companies or governments make human-centered data sets public, they put the data through significant processes to anonymize it. The intention is that no matter how inane the data seems, it should not be able to be traced back to the individuals it concerns. However, there are known mathematical issues with this process that make it difficult. Seemingly well-anonymized data can be, with the right techniques, made unanonymous by comparing one data set with other publicly available data sets.

A well-known example of this undoing of anonymity involves the company Netflix, which offers online streaming of TV shows and films. In 2006, Netflix released millions of anonymized movie ratings tied to specific users and offered the Big Data community a challenge: significantly improve their system of recommending movies to users with the newly-released dataset for a large monetary prize.[6] In just over a year, researchers with new mathematical methods

[6] Bennett, James, and Stan Lanning. "The netflix prize." *Proceedings of KDD cup and workshop*. Vol. 2007. 2007.

were able to de-anonymize the data by comparing "anonymous" user ratings from the Netflix dataset to public movie ratings on IMDb, another website that allows users to rate films.[7] On the surface, being able to tie movie ratings back to specific users might not seem particularly detrimental, but it had extremely negative consequences. When the non-anonymous ties between users and movies were was published, some users' personal information like political leaning and sexual orientation were revealed through the movies they had rated highly. Netflix faced invasion of privacy suits and serious backlash.[8]

This problem is not unique to Netflix. Any data set that is to be released publicly needs to be painstakingly altered by experts to ensure that privacy is maintained, especially in the field of development where data can be more personal. There are mathematical methods to prevent these kinds of mistakes, and they must be considered seriously and heavily throughout the process of using public data.

Development-Specific Big Data Criticisms

---

[7] Narayanan, Arvind, and Vitaly Shmatikov. "How to break anonymity of the netflix prize dataset." *arXiv preprint cs/0610105* (2006).

[8] Singel, Ryan. "Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims." *Wired*. Conde Nast, 17 Dec. 2009. Web. <https://www.wired.com/2009/12/netflix-privacy-lawsuit/>.

Self-Selection Bias

A main criticism of Big Data projects within development is that in the context of developing countries, many of the available, large datasets do not fully represent the populations researchers try to apply them to. Wherever technology is needed on the ground to collect the data, which is often, this problem is relevant. For example, if a project involves analyzing internet search terms, in a developing country context access to the internet may not be widespread. There may be and typically are disparities in who is able to access technology across class, gender, age, and other social boundaries where infrastructure is lower.

This idea has been well-demonstrated specifically for mobile phone users but is applicable elsewhere. As mobile phones are one of the most widely-proliferated technologies in less developed countries, many researchers choose to use mobile phone data for Big Data analysis. However, researchers have shown that mobile phone users are not always a representative set of their communities. Women, children, the elderly, the poor, and those with less education are less likely to be users of cell phones in Rwanda, for instance.[9] In other words,

[9] Blumenstock, Joshua, and Nathan Eagle. "Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda." *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. ACM, 2010.

the system benefits wealthy, educated men, and thus creates less representative data sets.

While the bias of Big Data sets within infrastructurally weaker countries is an important consideration for researchers to make when using data in this context, its existence should not necessarily rule out the usefulness of Big Data in this context altogether. Though the challenges are somewhat more unique within development, bias inherent in data is truly a universal data problem. Users of data anywhere should always take potential bias from the data set into consideration, and from reviewing the literature most all primary research within development takes this issue seriously and presents it up front. As long as researchers are aware of this obstacle, they can and do make efforts to work around it and are careful to not generalize results too widely.

Big Data methods cannot necessarily produce claims that are as specific as smaller data methods can, but this certainly does not make Big Data methods useless. The research itself has shown that the projections producible from less representative Big Data sets can have high levels of correlation with "official" results that were painstakingly made representative of local communities. As a result of this accuracy, Big Data could even be a semi-solution to the problem of not having representative data in the first place if analysts take these deficiencies into account properly. Self-selection bias in this field is critical to be aware of, but should not dismantle the usefulness of the results altogether, and is being validly overcome in practice.

Cultural Relevancy

A second problem more specific to the application of Big Data in development, and to applications in social sciences in general, is that if the dataset in use is being produced from the interaction between people and technology, it is crucial to understand the specifics behind how those people view the technology and use it. There exist various cultural norms and practices globally when it comes to technology use that if misinterpreted could result in inaccurate conclusions in Big Data for development projects. For example, a particular country may have different norms about mobile phone calls. To be more specific, if attempting to create a graph of the social networks in a community using mobile phone data, the researcher must understand how people in that community tend to view mobile phones, phone calls, and text messages. Do they think a phone call or a text message is more personal? Do they think it is rude to have conversation over the phone in public and will they choose to not call instead? Do they tend to trust mobile phone companies to keep their conversations private?

All of these considerations and more, depending on the research questions being asked, will vary from context to context and could affect conclusions. The same statistical model that worked in one setting on mobile phone data may not work in other settings. When applying results in different cultural settings, there is usually no one-size-fits-all model. Furthermore, cultures are constantly changing. The way communities interact with technology over time is not set in stone, and statistical models should not be either. Approaches must be constantly adjusted to account for cultural shifts in the way people interact with the world. Similar to the above discussion of self-selection bias, as long as practitioners are aware of this problem they can do something about it and make adjustments. Big Data for development projects that approach these nuances thoughtfully can sidestep the pitfalls associated with misinterpreting culturally-specific data, and again, most deal with this criticism explicitly.

Research Questions and Methodology

The main critiques of Big Data projects in general and of applications specifically within development and humanitarian aid do not appear to overwhelm Big Data's overall potential to create actionable insights. There looks to be a space for practitioners to apply Big Data methods to longer-term development problems like poverty or food security as well as shorter-term humanitarian aid issues like natural disaster recovery or disease outbreak management. What effort has been made in this space? Can the use of Big Data in international development and humanitarian aid projects create impact for the intended beneficiaries? What kinds of benefits and issues are development practitioners likely to encounter when using Big Data? This thesis aims to answer these questions.

Within the following four chapters, the literature on the applications of Big Data for development is distilled into three categories. Specific projects in each of these categories are analyzed to determine if the project had potential for impact. As few research papers have directly implemented their results, an impactful project for the extent of this paper is defined as a project that through its use of data created a solution to a problem that objectively improved upon the traditional method. If it was faster, more accurate, more encompassing, or in more general terms, would help development practitioners make smarter decisions, the project is determined to have potential for impact. If the results proved enormously difficult to assemble, were simply inaccurate, or did not improve upon the traditional method for solving the associated problem, the project's potential for impact dissipates.

Over one hundred related publications were sorted and analyzed with a focus on primary, academic research papers to distill the Big Data for development literature into distinct sets. Three types were devised based on temporal characteristics of the specific development or humanitarian aid problem involved. Each of the following three chapters represents an analysis of one of the three devised categories.

The first category, called Type I, encompasses ongoing development issues, or problems that are continuously occurring where the trend is relatively stable over time. An example of a Type I development problem, covered in Chapter 2, is that of ethnic segregation, as researchers are mostly interested in measuring it and assessing gradual change. Type II intermittent development problems are similar in that the issues are usually continuously occurring to some extent, but periodic intensity of the issue is what causes it to be a development problem. Chapter 3 covers Type II dilemmas, and includes discussions on multiple disease outbreak applications. The third category, or Type III, is discussed in Chapter 4, which focuses on sudden and typically non-repeating crises, an example being an earthquake. Distilling the literature into categories based on the type of development problem allowed for trends across each of the three categories to be illuminated. After discovering what potential for impact Big Data has for ongoing, intermittent, and sudden development problems, the fifth and final chapter discusses what barriers exist to utilizing these technologies to their fullest extent, and offers suggestions for the further incorporation of Big Data into the fields of development and humanitarian aid.

Chapter 2 | **Type I: Ongoing Issues**

The first category of development problems, also referred to as "Type I" within the scope of this paper, involves issues that are continuously occurring. They are ongoing problems where the dilemma is getting measurably better or worse over time, regularly on a scale as short as a single day. When Big Data is applied to Type I, ongoing issues, the datasets do not typically display recurring, abrupt changes or spikes of activity over time. What is interesting about the data in Type I projects is not the existence of defined events but long-term trends.

There are many different applications of Big Data to Type I issues within the world of international development and humanitarian aid. Example issues, which again are not characterized by consistently repetitive events but rather by longer-term patterns, are challenges like ethnic tension in local communities, the process of radicalization related to terrorism, and global malaria infections. Each of these problems disproportionately affect poorer global communities[10], and can create barriers to improving quality of life. Thus, the development community has for some time attempted to solve or alleviate each of them to varying degrees of success with a number of different methods.

---

[10] Stevens, Philip. "Diseases of Poverty and the 10/90 Gap." *Diseases of Poverty and the 10/90 Gap* (n.d.): n. pag.*International Policy Network*. World Health Organization. Web.

However, these problems are not solved, as evidenced throughout the rest of this chapter. Big Data is one approach that could be a piece of the puzzle. This chapter will explore each of the above ongoing development issues and what kinds of Big Data applications are working to help solve them. It will discuss whether the Big Data solutions significantly improve upon the traditional solutions, and tease out what is particular about the approach and solutions to Type I Big Data projects.

Application: Ethnic Segregation

Measuring Ethnic Segregation Using Mobile Phones

Data from the World Health Organization indicates that areas with high levels of ethnic segregation are more likely to experience incidences of violence, and that segregation is directly tied into a country's future prospects for development. Development institutions are thus interested in the reduction of ethnic segregation to further the goal of improving global well-being. Traditionally, efforts to reduce ethnic segregation are comprised of in-person systems, such as programs to encourage relationship building and power sharing across social boundaries in cities with noticeable ethnic tension[11]. However, it can

[11] Burg, Steven L. "Preventing ethnic conflict: Macedonia and the pluralist paradigm." *Online: http://www. wilsoncenter. org/index. cfm* (1997).

be difficult to determine what the best approach is as it is difficult to measure changes in ethnic tension. Traditional methods are slow and not comprehensive. Surveys are time consuming, could introduce bias in responses, and might require volunteers on the ground in dangerous environments.[12] Relying on official reports only slightly faster, but is not comprehensive. Recently, Big Data has been applied to the issue of measuring ethnic segregation.

A team of researchers, led by Joshua Blumenstock and Lauren Fratamico, attempted to generate a framework that could accurately measure ethnic segregation using anonymized data from a phone network.[13] The data they utilized included the phone's operating language, call locations, and call records indicating length of call and who the receiving phone user was. The authors created a statistical model using this data that estimates the extent to which members of two different ethnicity groups interact. The ethnicity of the phone users was assigned using the operating language of the phone, which turned out to be an accurate enough indicator based on initial surveys. Indicators of closeness resulted from measuring who called who, for how long, and from where. While the authors

---

[12] Smith, Tom W. "Measuring Racial and Ethnic Segregation." National Opinion Research Center, University of Chicago, Apr. 2002. Web. <http://gss.norc.org/Documents/reports/methodological-reports/MR096.pdf>.

[13] Blumenstock, Joshua, and Lauren Fratamico. "Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data." *Proceedings of the 4th Annual Symposium on Computing for Development*. ACM, 2013.

recognize the limits of their framework in their paper, they believe through various validation methods that they have created a way to measure levels of ethnic segregation in a real-time, fine-grained way. Their model can be applied across neighborhoods and produce results back within the time frame of a single day.

If this data under further research continues to prove accurate, it could provide officials with real-time analytics on the potential for ethnic violence due to anomaly detection features within the system. Ethnic segregation is tied into violence and development, thus the ability to respond to issues of segregation is a crucial battle in overcoming violence as a larger development issue. By giving development practitioners a potential way to measure ethnic segregation, the researchers are enabling a number of positive changes that improve upon the traditional methods.

First, development officials would be able to measure what effect their programs are having in near real-time with this method. This would allow them to thoughtfully adjust their methods and make them more responsive to the community's shifting needs. Second, this method does not require boots on the ground or extensive surveys and interviews in order to understand the state of social group relations in a country. Compared to having community members take physical surveys, the Big Data method improves comprehensiveness in areas that have mobile phone networks or are too dangerous to put researchers on the ground. Third, this method provides information on a daily basis where in-person analyses of the same issues from surveys may take days or weeks to compile. The Big Data solution predominantly improved the comprehensiveness of the traditional approach, but also improved its speed and adjustability. If the method is further verified, it could serve as a new, impactful way to monitor and help reduce ethnic tension.

Application: Malaria

Tracking Populations to Understand the Spread of Malaria Using Mobile Phones

While individual cases of and mortality rates from malaria have been dropping over the last few years, the disease still remains a prominent global health issue. Over 90% of both the cases and deaths from malaria occur in Sub-Saharan Africa.[14] In total, around the globe in 2015, over 400,000 people died from the disease, which is both preventable and treatable. Malaria is caused by the transmission of parasites from mosquitoes to humans following a bite from the insect, or from the transmission of parasites from infected persons to others.

[14] "Fact Sheet about Malaria." World Health Organization. World Health Organization, n.d. Web.

In areas of frequent malaria transmission, “almost everyone” carries the parasite but will lack symptoms, making these unknowingly infected people primary transmitters of the disease.[15] This becomes a significant health issue when these asymptomatic carriers travel, bringing malaria to new populations who lack the partial immunity the carriers possess. As transportation within and between countries in Sub-Saharan Africa becomes more common and the paths more complex, estimating parasite routes to prevent the spread of malaria becomes more difficult. Some researchers have indicated that “although many qualitative surveys have explored the impacts of travel and transportation on health, economics, and development in Africa, there is a huge deficit of quantitative data on individual mobility from these regions.”[16]

In an attempt to solve this problem, these same researchers undertook a project utilizing large amounts mobile phone data to try to better measure the movement of populations, and specifically how this movement applies to the issue of malaria. Mobile phone companies in Kenya provided the team with anonymized call record data which allowed them to determine the location of phone calls within a few kilometers. They used this data to track individuals’ movement over time,

[15] Buckee, Caroline O., et al. "Mobile Phones and Malaria: Modeling Human and Parasite Travel." *Travel Medicine and Infectious Disease*, vol. 11, no. 1, 2013., pp. 15-22 doi:http://dx.doi.org.proxy.lib.umich.edu/10.1016/j.tmaid.2012.12.003.

creating patterns. Resulting maps of movement patterns were combined with maps of malaria cases, mobile phone towers, and population distribution. From analysis of these mas, the researchers concluded that "it is possible to pinpoint specific settlements with high risks of imported malaria and generate maps that show how the parasite might be traveling around a country within human carriers."[16]

Surveys could have provided the researchers with this information, but the process of gathering that data is slow, and it requires significant in-person human effort across much distance. Therefore, the primary benefit Big Data brought to malaria prevention effort was an improvement to comprehensiveness, as it was easier to generalize their model to an entire country than it would have been to do the same with survey methods. Speed was another benefit, but it was less critical in this scenario. Survey-style disease surveillance is still useful, but the World Health Organization notes that "currently many countries with a high burden of malaria have weak surveillance systems and are not in a position to assess disease distribution and trends, making it difficult to optimize responses and respond to outbreaks." In these cases, the Big Data method might be able to provide information on population movement that is not available at all through traditional approaches. Overall, a Big Data approach created improvements most directly to comprehensiveness when applied to the issue of malaria prevention.

Application: Radicalization

Detecting ISIS Sympathizers Using Social Media

Some Big Data projects have used posts from Twitter in their approaches to solve Type I development issues. One example is the use of statistical analysis on Tweets to try to detect ISIS sympathizers, ultimately with the goal of preventing violence.[16] ISIS is somewhat unique in the global sphere today as a terrorist group for its heavy use of social media to recruit potential members. Twitter has been fighting an uphill battle to try to prevent inspiration for this type of violence from taking place on their website. However, it has been difficult for the company to detect which of their users are indicating support for ISIS's radicalizing propaganda. They sometimes rely on user-to-user reporting mechanisms or trust-building programs[17] which are slow and do not reach all cases of radicalization occurring on their platform.

Big Data techniques were recently applied to this dilemma to try to automatically detect Tweets containing jihadists messages from a sea of posts on

---

[16] Ashcroft, Michael, et al. "Detecting jihadist messages on twitter." *Intelligence and Security Informatics Conference (EISIC), 2015 European*. IEEE, 2015.

[17] Bizina, Margarita, and David H. Gray. "Radicalization of Youth as a Growing Concern for Counter-Terrorism Policy." *Global Security Studies* 5.1 (2014): 72-79.

the site so that Twitter may choose to take action in response. The researchers were able to achieve high statistical accuracy with their categorization model, outperforming previous algorithms for detecting these kinds of messages which lacked both speed and accuracy, and their new method was certainly faster than user-to-user reporting mechanisms. Their Big Data approach covers more users of the site than previously could be managed, and for an issue like radicalization which takes place slowly over time, this comprehensiveness was critical. The algorithm also had the added benefit of speed over the reporting model, but is likely similar in speed to other less accurate algorithms. While the researchers indicate that this tool should be used to support existing methods and not to replace them, they have provided a valuable and more comprehensive way to detect radicalization in near-real-time. Therefore, development practitioners may find this useful when attempting to prevent future violence and to improve global well-being.

Conclusion

There are many ways Big Data has been used to solve Type I, ongoing issues such as malaria, radicalization, and ethnic segregation. Strategies like survey-based disease surveillance systems for malaria, identifying potentially radicalized social media users through user-to-user reporting mechanisms, and extensive surveys to determine risk or level of ethnic tension are all methods that have been used in development projects to try to alleviate these issues. As demonstrated above, Big Data techniques can be applied to each of these issues and in most cases do seem to improve upon the aforementioned traditional methods.

The primary benefit identified in Type I projects was improved comprehensiveness. When the Big Data method's results were compared with traditional methods, typically the Big Data method applied to many more people than the traditional method alone could cover. However, there are also risks that one should be particularly aware of when dealing with Big Data approaches to Type I problems. Of all the criticisms aimed at Big Data covered in Chapter 1, the most relevant criticism for ongoing issues is self-selection bias. As comprehensiveness is the primary goal and benefit from Big Data approaches in Type I, it is important that populations are accurately represented. However, disparities in access to technology can bias data sets, especially in less developed countries, making comprehensiveness difficult. If a project tackling a Type I issue does not adequately address these access differences, the measurements and categorizations produced will likely be inaccurate. There are ways around this obstacle. Researchers can choose to compare their results to official, representative data to confirm that it still reflects all populations, or they can determine through careful study if the experiences of underrepresented populations are different enough from those represented through technology interactions to warrant a different approach. Practitioners must consider this issue thoughtfully especially when attempting to solve Type I problems with Big Data.

Chapter 3 | **Type II: Intermittent Issues**

The second type of development dilemma, also referred to as “Type II” within the scope of this paper, involves development problems that are also known and ongoing, but where the data is characterized by abrupt, repetitive spikes. They are considered intermittent issues because the primary interest of development practitioners is the timing intensity of these defined, recurring events.

Challenges like preventing, defining, and treating outbreaks of the flu, one of the globe’s most common infections[18], and similar challenges for dengue, which present flu-like symptoms but has a different treatment timeline, are examples of Type II intermittent issues. These health problems might be continuously ongoing to some extent, but what is most important about the data involved is that there exists recurring, severe events. There would likely still be efforts to prevent and treat the flu even flu cases were steady over time (i.e. more similar to data in Type I issues), but it is the massive spike in cases that occurs during each flu season that is the distinguishing and central obstacle in managing the infection. It is this pattern of events over time that distinguish Type II issues from problems considered Type I.

[18] "Flu Symptoms & Complications." *Centers for Disease Control and Prevention*. N.p., 23 May 2016. Web. <https://www.cdc.gov/flu/about/disease/complications.htm>.

Development practitioners are often interested in these problems as they disproportionately affect poorer global communities[19], creating barriers to improving general quality of life. The development community has attempted to alleviate the issues caused by both the flu and dengue to varying degrees of success. Some countries have put disease surveillance systems in place involving in-person reporting mechanisms that measure the number and location of cases of various diseases and infections through hospital admissions.[20]

Big Data is one tool that has been applied in a research setting to both the flu and dengue in an attempt to improve tracking and prevention. This chapter will explore Big Data Type II applications helping to solve intermittent development problems, discuss whether the Big Data solutions significantly improve upon the traditional solutions that exist today, and identify what is particular about the approach and solutions to Type II Big Data projects.

Application: Flu

Calculating Flu Cases using Internet Search Terms

---

[19] Stevens, Philip. "Diseases of Poverty and the 10/90 Gap." *Diseases of Poverty and the 10/90 Gap* (n.d.): n. pag.*International Policy Network*. World Health Organization. Web.

[20] "Communicable Disease Surveillance and Response Systems." *PHLS* (n.d.): 32. World Health Organization. Web. <http://www.who.int/csr/resources/publications/surveillance/WHO_CDS_EPR_LYO_2006_2.pdf>.

Google has a global monopoly on search engine users, managing trillions of queries per year.[21] Many websites on the internet contain information about health problems that users can easily look up through a search engine like Google. If a user has a sore throat, runny nose, and a dry cough, symptoms common with the flu, it is much easier and faster for them to type these symptoms into Google than it is for them to visit a health professional. The result of this convenience is that the company is the recipient of a large quantity of symptom-related queries.

In 2008, Google decided to start doing something with this data: the company starting testing its ability to predict incidences of the flu using historical queries from users that match the appropriate symptoms in a project called Google Flu Trends. The use of internet search terms as a source of Big Data was extremely new at the time. Google's predictions were surprisingly statistically accurate when compared with government-collected flu data from the same historical period of time.[22] The company quickly expanded the project's scope and started using real time queries to make predictions, expanding into over 30 countries while continually corroborating its results with official data.

---

[21] "Search Engine Market Share." NetMarketShare, n.d. Web.

[22] Cook, Samantha, et al. "Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic." *PloS one* 6.8 (2011): e23610.

If Google Flu Trends was recreating the same data already collected by national disease surveillance agencies, what was the point of going through all the trouble of collecting the search terms and creating their algorithm? As Google states: "It turns out that traditional flu surveillance systems take 1-2 weeks to collect and release surveillance data, but Google search queries can be automatically counted very quickly. By making our flu estimates available each day, Google Flu Trends may provide an early-warning system for outbreaks of influenza. For epidemiologists, this is an exciting development, because early detection of a disease outbreak can reduce the number of people affected"[23]

However, Google Flu Trends is no longer in action. While it remained steadily accurate throughout 2011, in the summer of 2012, the tool abruptly stopped producing accurate numbers, with predictions rocketeering far past what was reported by government institutions.[24] What happened? Some researchers suspect that search behavior dramatically changed as the result of an unexpected summer outbreak, thus undermining the cultural logic supporting the algorithm and resulting in predictions that were vastly too high. The algorithm was made on

---

[23] "Tracking Flu Trends." *Official Google Blog*. N.p., 11 Nov. 2008. Web. <https://googleblog.blogspot.co.il/2008/11/tracking-flu-trends.html>.

[24] Lazer, David, et al. "The parable of Google flu: traps in big data analysis." *Science* 343.6176 (2014): 1203-1205.

assumptions about people's search behavior, and the search behavior changed while the algorithm did not adjust. While this specific project was shut down, Google still walked away with some successes. Google Flu Trends was the first tool to demonstrate that internet search terms held promise for disease prediction, as their accuracy in the first few years suggested.

In its prime, Google Flu Trends operated in almost real-time and was able to provide accurate flue predictions in countries which had little infrastructure in disease surveillance. It was primarily faster in ways that the traditional methods could not have been, providing evidence of potential for impact. However, it failed to adjust to changing social behavior and therefore could not maintain its cultural relevancy.

Application: Dengue

Calculating Dengue Cases Using Internet Search Terms

Dengue is an infectious disease spread through mosquitoes, and populations in tropical environments are most at risk.[25] The infection bears a few commonalities with the flu, causing flu-like symptoms and similar

[25] "Dengue and Severe Dengue." *World Health Organization*. World Health Organization, n.d. Web. <http://www.who.int/mediacentre/factsheets/fs117/en/>.

seasonally-timed outbreaks.[26] If not detected early enough, dengue can proceed to a particularly severe stage, after which there is no consistently successful cure, therefore it is vital to track dengue symptoms to be alerted as early as possible to infections to prevent loss of life.

Following the successful tracking of the flu with internet search terms, some researchers applied a similar model in an attempt to track dengue. One particular project used a sample of historical, anonymized Google search data including queries relating to dengue. They created a model that attempted to predict dengue outbreaks based on these queries, verifying their results with open data from various governments surveillance systems.[26] In both countries studied, Singapore and Thailand, correlations between the team's predictions and the official government data were high: between 0.8 and 0.9 across the board. Similar to the aforementioned study on the flu, it was deemed possible to create an algorithm to determine cases of the disease without needing any actual reports of the disease.

The two country-based case studies chosen in this research paper have highly contrasting levels of infrastructure. Singapore has a rather sophisticated disease surveillance system where much of Thailand is lacking. However, the authors note that the predictions made by their models could be useful in both

---

[26] Althouse BM, Ng YY, Cummings DAT (2011) Prediction of Dengue Incidence Using Search Query Surveillance. PLoS Negl Trop Dis 5(8): e1258. doi:10.1371/journal.pntd.0001258

settings. In Singapore, where government data is available in only a week's time, they suggest their Big Data conclusions be used as a supplement to speed along and confirm the traditional process. For Thailand, however, they note "there are significant delays in the reporting of cases from many areas of the country. Our model may give significant improvements in settings with significant delays. It is conceivable that some dengue-endemic settings in South and Southeast Asia may have significant internet use before surveillance systems are developed and thus an internet search term-based model may be a proxy for routine surveillance in these settings."

Therefore, while the authors still recommended the use of their conclusions in countries with more seasoned disease surveillance systems, they were particularly adamant about the potential impact for countries which contain little to no surveillance infrastructure yet which have internet access. By taking advantage of the existence of large numbers of internet search queries, they were primarily able to find important trends faster than they could previously. In some but not all scenarios, they could produce results where there was previously little to no government data to be found at all. This disease-related information can help governments and organizations make an impact in disease prevention and treatment through the allocation of resources to the right place at the right time.

Calculating Dengue Cases Using Mobile Phones

While certainly promising, internet search terms are not the only way Big Data technologists have found to track the spread and movement of infectious diseases. When it comes specifically to dengue, a detailed research project was carried out in Pakistan to discover whether large volumes of phone call data could predict dengue incidences with any accuracy.[27]

The research team's specific source of data was phone calls from an emergency health line. Through analyzing dengue-related symptom calls and questions over time, they were able to create a model based on the calls that attempted to calculate the general intensity of dengue cases. The researchers adjusted their model every few weeks as new sets of incoming data arrived to try to ensure that the model changed as call behavior evolved. Their model predicted dengue incidences with a 0.85 correlation compared to the official government data, but provided it two to three weeks earlier.

This mobile phone dengue monitoring system created by the researchers was actually adopted by Pakistan's government in Lahore. A website was created that allows government officials to receive visualized dengue predictions two weeks in

[27] Rehman, Nabeel Abdur, et al. "Fine-grained dengue forecasting using telephone triage services." *Science Advances* 2.7 (2016): e1501215.

advance of their official data. The team's strategies correlated with a decrease in Lahore's number of dengue-infected patients from over 20,000 in 2011 to around 1,500 in 2013. While no direct causation can be implied, the predictions created from the large dataset of health-related phone data proved to be highly accurate and much faster than the previous solution while remaining comparatively technologically simple on the ground.

Conclusion

Intermittent issues like disease outbreaks have been tackled by development practitioners for a long time. Traditional strategies typically involve government-backed disease surveillance systems. However, data from these systems take anywhere from one to three weeks after cases are reported to be collected and disseminated for use. Speed is especially critical when it comes to preventing disease mortality as some diseases, like dengue discussed above, need to be addressed early for treatment to be effective. The few days to weeks that the Big Data solutions saved on time could mean life or death for people suffering from such diseases. This increased speed is the primary benefit seen in most Type II Big Data for development projects. If there is already a traditional method in place, like a disease surveillance system, it is typically already somewhat comprehensive, though a Big Data approach may still improve comprehensiveness when it can. If there is not yet a traditional method in place, such as in some countries with less infrastructure, the Big Data method can provide data that did not exist at all previously and revolutionize comprehensiveness, though this is less common.

The one challenge most associated with Type II projects is the cultural relevancy of the model as discussed in Chapter 1. This is in part because the central benefit of Type II is speed. When data becomes available, it needs to be analysed as quickly as possible for the benefits of an early warning system to be maximized. If practitioners are focusing on boosting speed, there is less time available to consider if the data looks or behaves differently than data from previous samples. This can result in lost accuracy as the behavior of people changes but the model does not adjust accordingly. Culture is gradually but constantly changing, therefore a culturally-based statistical model must be constantly adjusted to accommodate these shifts. Making adjustments can be made additionally difficult due to simply the nature of Type II issues as defined events, offering researchers less opportunities to notice discrepancies in data. Those spearheading Big Data for development projects, but especially projects that are categorized as Type II focused on speed, must keep this in mind and readjust models frequently to maintain the highest possible level of impact.

Chapter 4 | **Type III: Sudden Crises**

The final category of development dilemmas, also referred to as "Type III" within the scope of this paper, refers to sudden, humanitarian crises that are not predictably repetitive. Examples of Type III sudden disasters are earthquakes, tsunamis, violence, and non-repetitive disease outbreaks. The data typically used in Type III projects is more wide ranging than the previous two types, as the projects usually take place in emergency contexts when there is less prepare data sets in advance, and practitioners do not have the luxury of choosing the location of study. Due to less flexibility, whatever locally available data exists in these contexts is potentially useful.

Type III issues are often categorized as humanitarian aid-focused rather than development-focused because the aims of projects in this field are shorter-term. However, longer-term development aims are inevitably intertwined within these shorter-term humanitarian aid issues as sudden crises disproportionately impact poorer global communities long after the initial emergency subsides[28] [29]. Therefore

---

[28] Hallegatte, Stephane, et al. *Unbreakable: Building the Resilience of the Poor in the Face of Natural Disasters*. World Bank Publications, 2016.

[29] Sakai, Yoko, et al. "Do Natural Disasters Affect the Poor Disproportionately? Price Change and Welfare Impact in the Aftermath of Typhoon Milenyo in the Rural Philippines." *World Development* (2017).

while short-term needs are characteristic of Type III, both the development and humanitarian aid communities make continual efforts to respond to these types of crises in order to create positive change that impacts both the short- and long-term.

While many humanitarian efforts targeted at global crises place more of an emphasis on preventative measures, responses are still critical after a crisis hits. Typically, they focus on effectively managing the resulting population displacement and resource availability issues caused by new environmental or societal dangers. This involves the coordination of resources between many different national and foreign organizations, specifically items like food, health supplies, and emergency kits, and the creation of shelters or water purification systems where deemed necessary.[30]

Despite these post-disaster humanitarian efforts, the resulting population and resource stressors occurring as a direct result of crises are continuing to take a toll on humanity. In the last decade, as a direct result of only weather-related natural disasters, over 4 billion people were in need of immediate aid and over 600,000 people ultimately perished.[31] This chapter will explore how Big Data has been

---

[30] "USAID's Office of U.S. Foreign Disaster Assistance Fact Sheet." *USAID* (n.d.): n. pag. 2017. Web. <https://www.usaid.gov/sites/default/files/documents/1866/ofda_fact_sheet_02-09-2017.pdf>.

[31] CRED, UNISDR. "The human cost of weather-related disasters, 1995–2015." *United Nations, Geneva* (2015).

applied to alleviate the suffering caused by sudden Type III issues. It will then discuss whether the Big Data solutions significantly improve upon the traditional solutions to solve the same issues, and identify what is particular about the approaches and solutions to Type III Big Data projects.

Application: Earthquake

Tracking Population Displacement with Mobile Phones

The 2010 earthquake in Haiti resulted in the deaths of over 200,000 people and impacted more than 3 million.[32] The quake struck near Port-au-Prince, Haiti's densely populated capital city. There were few building codes in place at the time, regional infrastructure was low, and many people were living in informal concrete settlements. These factors exacerbated the disaster's impact through increased susceptibility to structural damage. It is estimated that the earthquake left over one million Haitians homeless and in need of aid.

Generally, post-disaster, it is critical for on-the-ground aid organizations to facilitate the timely and efficient distribution of resources to those who need them most. These resources were imperative for Haitians in affected areas in the days and weeks that followed the earthquake. However, effective coordination of

---

[32] Pallardy, Richard. "Haiti Earthquake of 2010." Encyclopædia Britannica. Encyclopædia Britannica, Inc., 18 Aug. 2016. Web. <https://www.britannica.com/event/Haiti-earthquake-of-2010>.

disaster response resources can be difficult to achieve due to the fact that disasters tend to cause chaotic and widespread population movements. Often, population distribution information is gathered through volunteers who rely on "eyewitness accounts, manual counting of people, registration of people in camps, or satellite or aerial images of shelters or changes in vegetation."[33] If there were a more efficient method, the most desperate victims of disasters could receive aid faster at the time when they need it most.

One possible alternative to tracking population movements is the use of location data collected by mobile phone networks. When mobile phone users make calls, the network provider records the location of each call. A team of researchers worked with the largest mobile phone company in Haiti, Digicel, to determine if historic logs of phone call locations could accurately uncover grander and more generalized movements of people in the post-earthquake context in 2010. By analyzing the locations of calls of almost 2 million unique Digicel customers, the team produced maps of population movements within Haiti that closely matched reliable though slow data collected by the UN well after the initial disaster subsided. Furthermore, the initial population estimates that were circulated

[33] Bengtsson, Linus, et al. "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti." PLoS Med 8.8 (2011): e1001083.

throughout disaster response organizations at the time of the crisis, used heavily to coordinate resources, were found to be less accurate than both the UN data and the information procured from the mobile phone networks.

The authors favorably suggest that tracking population movements in crisis settings could be “revolutionized in areas with high mobile phone coverage” noting that “86% of the world’s population lives under mobile cellular network coverage.”[33] Overall, the use of the data from the Digicel network in Haiti provided more accurate estimations of population densities than the information circulated amongst aid groups at the time of the earthquake. The cell network method could have provided this population information in near-real-time compared to the days and weeks required to gather and communicate results from surveys and manual counting of people.. In addition to being faster, the results were more also more comprehensive. Manual population counting efforts only took place primarily in ports and cities. However, the mobile networks in Haiti cover 90% of inhabited areas, and recovering the cell network information required no teams of volunteers on the ground in potentially dangerous crisis settings. The benefits witnessed to both speed and comprehensiveness foreshadow a future potential for impact if aid organizations incorporate this Big Data method into their arsenal.

Application: Cholera Outbreak

Tracking Cholera Outbreak Location and Intensity with Mobile Phones and Social Media

The same team of researchers who utilized data from the Digicel network in Haiti after the 2010 earthquake applied a similar method to a different kind of humanitarian crisis in Haiti: a cholera outbreak. Cholera is an acute infection caused most often by drinking contaminated water.[34] If left untreated, in some persons it can cause death within a matter of hours. It can be challenging to track and prevent because the majority of infected persons show no symptoms, but can pass on the infection for up to 10 days. Therefore it is imperative that response teams understand who has passed through cholera-infected areas and where they are going in order to respond to and prevent current and future infections. The researchers' findings indicated that again, accurate population movement estimates could be generated and verified, accurate down to specific hours of the day, which could lead to "important improvements in the allocation of relief supplies and the quality of needs assessment surveys."

A different Big Data project was also undertaken for the same Haiti cholera outbreak instead utilizing social media to try to determine outbreak intensity over

[34] "Cholera." World Health Organization. World Health Organization, Oct. 2016. Web. <http://www.who.int/mediacentre/factsheets/fs107/en/>.

time. Both Tweets and data pulled from an informal media source called HealthMap were analyzed to try to determine if they could be used to predict cholera outbreak intensity. The results were compared to official data collected over a period of weeks from the Haitian Ministry of Public Health. The authors ended up discovering a "good correlation" between the official data and the predictions made from their two social media sources.[35] By analyzing large amounts of posts from Twitter and HealthMap, they could provide information about the intensity of the cholera outbreak within a number of hours compared to the weeks it took Haiti's health ministry to compile the data, government results "which are [only] available after delays incurred in the traditional chain-of command structure of public health. The use of electronic sources can also facilitate finer temporal resolution than more traditional data streams; often at the level of single days or better."

However, important limitations hindered the usefulness of this application of social media. Analytics provided through the dissemination of the selected social media sources on the intensity of the cholera outbreak lost accuracy as more time passed after the first spike in cases. The authors attribute this issue primarily to the

[35] Chunara, Rumi, Jason R. Andrews, and John S. Brownstein. "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak." The American journal of tropical medicine and hygiene 86.1 (2012): 39-45.

fact that after the shock of an initial news story breaks, online social media users are less likely to report further incidences. While long-term tracking of disease outbreaks might not benefit, in the short-term right after an outbreak strikes, social media sources could still provide earlier tracking of infections. These limitations are acknowledged in full by the authors who note, “The methods here are primarily useful for evaluating the relationship between informal and official data streams during periods of high disease transmission activity, which commonly occurs at the beginning of an outbreak.” Therefore, while the results do not promise an all-encompassing method, they do improve upon a piece of the disaster response puzzle specific to early monitoring of disease outbreaks by improving the speed at which organizations could be notified.

Application: Floods

Flood Detection and Assessment with Social Media and Satellite Signals

According to a comprehensive study on the global impact of floods over almost the last thirty years, floods are “the leading cause of natural disaster fatalities worldwide.”[36] Between 1980 and 2009, as a direct result of floods, almost

[36] Doocy S, Daniels A, Murray S, Kirsch TD. The Human Impact of Floods: a Historical Review of Events 1980-2009 and Systematic Literature Review. PLOS Currents Disasters. 2013 Apr 16 . Edition 1. doi: 10.1371/currents.dis.f4deb457904936b07c09daa98ee8171a.

3 billion people were negatively affected and over 500,000 perished. Globally, floods tend to disproportionately impact poorer communities. One of the reasons behind this unequal impact is geographic, explained well by the World Bank: “As the coastal cities of Africa and Asia expand, many of their poorest residents are being pushed to the edges of livable land and into the most dangerous zones for climate change. Their informal settlements cling to riverbanks and cluster in low-lying areas with poor drainage, few public services, and no protection from... flooding.”[37] Another reason for this disparity is simply that countries with less resources to spare have less prevention infrastructure and are also less able to adequately respond to victims’ needs after floods strike.

Post-flood, humanitarian disaster response teams need timely and accurate information concerning the location and severity of damage in order to appropriately respond to needs for food, water, sanitation, first aid, and shelter. Without this information, it is difficult to minimize mortality. Unfortunately, neither humanitarian organizations nor governmental institutions have a

---

[37] "What Climate Change Means for Africa, Asia and the Coastal Poor." World Bank. N.p., 19 June 2013. Web. <https://www.worldbank.org/en/news/feature/2013/06/19/what-climate-change-means-africa-asia-coastal-poor>.

widespread, fast, and accurate way to track areas of risk following floods.[38] Typically, weather forecasters work to locate potentially high-risk areas after a flood strikes, but according to Emily Niebuhr, a meteorologist with the World Food Programme, "Flood and rainfall information varies widely by country….Sometimes it's not available at all."[39] Apart from meteorologists, traditional methods include efforts taking place on the ground "through a network of field stations, employees, and volunteers, as well as through common news outlets, such as radio and television," however there is a non-insignificant delay between some number of hours to days until humanitarian organizations receive this information about current damage and near-future areas of risk after and during floods.[40]

A team of researchers in the Netherlands undertook a Big Data project to try to understand the effectiveness of using data from both satellites and social media as a way to solve this data disconnect. Framing their study around a number of

---

[38] Bengtsson, Linus, et al. "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti." PLoS Med 8.8 (2011): e1001083.

[39] Garner, Rob. "Satellite-Based Flood Monitoring Key to Disaster Response." NASA. NASA, 21 July 2015. Web. <https://www.nasa.gov/feature/goddard/satellite-based-flood-monitoring-central-to-relief-agencies-disaster-response>.

[40] Jongman, Brenden, et al. "Early flood detection for rapid humanitarian response: harnessing near real-time satellite and Twitter signals." ISPRS International Journal of Geo-Information 4.4 (2015): 2246-2266.

historical floods, to gather their data they used publicly available tools to gather past Tweets with related keywords, and pulled public satellite imagery from the experimental Global Flood Detection System (GFDS) which uses a form of microwave to measure inundation. Both of these data sources are near-real-time as Tweets are available within minutes of being posted, and data from the GFDS is available within twenty four hours. The researchers performed a comparative analysis of data from disaster response agencies, their selected Tweets, and the GFDS data to discover if their two non-conventional sources of social media data could live up to the accuracy of the official data and provide relevant, fast flood information. For each of their three data sources, they analyzed its capacity to map flooded areas, detect floods early, and build understanding of the disaster in context.

While they eventually conclude that both sources of data have "huge potential," their results were mixed.[37] In terms of creating maps of flood severity, the GFDS satellite data performed relatively well, providing maps similar to those produced by disaster relief organizations at the times of the floods studied. It could have done so within a day, much shorter than the several weeks it took relief organizations to produce severity maps, improving speed. Additionally, if a system like the GFDS were used consistently to create these maps, comprehensiveness would also improve as satellites can be positioned to image anywhere on Earth and avoid on-the-ground difficulties volunteers might encounter. However, the Twitter data performed poorly at spatially mapping the selected floods. The researchers recognize that most Tweets are providing second-hand information, and that social media users are concentrated in urban areas, which made it extremely difficult to determine what location flood-related Tweets were discussing.

In terms of early detection, in some cases Tweets did precede official reporting of flooding incidents by a number of days, and thus could have given organizations earlier notice, but this observation is not consistent across each flood studied. The satellite data's performance was similarly inconsistent at early detection. It depended heavily on the environment being monitored, as generally the system is much more accurate in dry, unirrigated areas and thus produced more positive early detection results in Pakistan than it did in the Philippines.

Overall, the researchers remain positive and hopeful about the future use of non-conventional data sources like Twitter and satellite signals to improve the humanitarian response to floods. There were cases in which analysis of the large data sources pulled from Twitter and the GFDS provided faster, relatively accurate information about the severity, location, and timing of a flood that could have provided a benefit to aid organizations. While the GFDS did produce useable maps much more quickly than the traditional method, the system generally has room to improve. It is a newer innovation in flood detection and has known technical flaws that prevent it from most accurately measuring inundation across contexts, and would require further improvement before being generally applied. Additionally, while social media proved somewhat useful at early detection, it was difficult to use Tweets for large-scale location-based tracking due to Twitter usage patterns causing difficulties in assigning locations. There appears to be promise being hindered both by difficulties with the technology itself and by data availability, though these early results do show improvements in both speed and comprehensiveness.

Conclusion

Humanitarian aid and disaster response groups need accurate information quickly in the aftermath of a sudden crisis like an earthquake, a cholera outbreak, or a flood. Information concerning the location of affected populations and the intensity of the crisis at hand is imperative in order to accomplish the effective distribution of resources to prevent further injury or death. This information is usually collected through surveys or in-person observations, which can take days to weeks to compile and is subject to a number of different biases inherent in the observations taken and the limited number of locations that can be covered. As demonstrated, careful use of Big Data sources like mobile phone records, social media posts, and satellite signals can be used to provide information on the distribution of people, risk, and need following Type III crises.

The primary benefits witnessed are highly dependent on the situation, but tend to be more diverse than those in Type I or II scenarios. Both improvements to speed and comprehensiveness of disaster information appears to be typical. The wider improvements witnessed are potentially due to the fact that in emergency contexts, the traditional methods are already resource- and time-strained, and the methods have had less opportunities to be tested in practice than those used in Type I or Type II scenarios, allowing generally for more room for improvement to be made by Big Data.

Generally, it does appear that applying Big Data to sudden crises is more challenging. Crises are unpredictable and might disrupt technology use. In the case of earthquakes, if an earthquake takes out mobile coverage as did happen temporarily in Haiti, cell phones could not be used to track populations. In terms of cholera monitoring, while mobile phone data performed well, social media data was simply not suited to the fine-grained type of responses that disaster response teams required. And finally, in the case of flooding, current satellite systems are not quite technologically up to the task of providing consistent results across different environments. The needs in the midst of sudden crises are typically more immediate and more widespread than the needs of beneficiaries suffering from ongoing or intermittent development problems. It is much more difficult, therefore, to produce an encompassing Big Data solution, and might lead more desperate practitioners to use a dataset not well-aligned to the situation at hand due to lack of alternatives, reducing accuracy. However, as the above case studies demonstrated, Big Data methods can provide comprehensive information faster and than the traditional method in some cases.

Chapter 5 | **Conclusion**

Overview of Findings

The two questions this thesis primarily sets out to answer are what effort has been made in the Big Data for development sphere, and whether or not these efforts have potential for impact. After much exploration, the literature surrounding Big Data for development was found to be fairly small, as discussed in Chapter 1. There is much buzz about Big Data's general potential and many critiques of its use. Some authors believe that the concerns, which encompass worry over inherent uncertainty in the statistical methods used, a running theme of dangerous overconfidence, issues of data privacy, unequal distribution of technology in less developed countries, and the challenge of interpreting results in the context of culture, overwhelm Big Data's potential and that it should not be applied to development problems. However, each of these concerns is briefly discussed in Chapter 1 to demonstrate that with the proper preparation and mindset, a Big Data project can produce valid results and overcome these obstacles for the most part.

Moving forward with the understanding that Big Data for development is potentially worthwhile, the primary research was disseminated. Compared to articles that praise or critique Big Data, significantly fewer examples exist of it actually being applied to a development issue. While few in number, only these types of primary, academic research papers were considered potential case studies. This was simply due to the fact that full knowledge of the methods and intended uses of projects was required, a certain level of detail found only in academic papers. This set of case studies was then broken up into three, manageable categories based on the temporal nature of the development problem involved: ongoing, intermittent, or sudden. At least three Big Data applications were selected as case studies in each category. The purpose of this design was to gain a general understanding of the types of applications of Big Data in development to date, to create several opportunities to assess potential for impact, and to discover if any themes existed within the created categories that might assist future practitioners. Within the exploration of the case studies, first each is demonstrated to be of significant importance in the field of development. Then, the results of using the current or traditional method of approaching the problem is compared to the results found using a Big Data solution. If Big Data created a faster, more comprehensive, or otherwise better way to help solve the problem, it is considered to have potential for impact in that context.

Chapter 2 covered applications of Big Data to Type I development problems, or ongoing issues. These are problems that, for the foreseeable future, do not necessarily have a beginning or an end, but are constantly occurring to some extent. Potentially, unusual spikes in data related to ongoing issues might exist, but these spikes are not of primary interest. It is knowledge of the general trend over time that interests those in development hoping to reduce the negative impacts of the problem at hand in order to improve global quality of life.

Specific to this thesis, the Type I problems discussed are ethnic segregation, malaria, and radicalization. Manual surveys and classification methods are often used to cover large populations in order to gain a sense of how these problems are operating. While the Big Data approaches to each of these issues showed multiple, consistent improvements from these traditional methods, the most important improvement from a development perspective was increased comprehensiveness. The Big Data method of measuring these issues reached more people more easily which is more important for Type I issues than speed, though this also saw a boost.

Self-selection bias is the most prominent pitfall of Type I Big Data projects. Comprehensiveness is central to a successful approach to a Type I problem, as the main point is gaining a general understanding of the state of the issue in the world for all populations. Yet, this can be a challenge due to the fact that people, especially in developing country contexts, do not interact equally with techology. If the dataset being used is produced from the interaction between people and technology, it is likely the dataset is missing information about women, children, and poorer families. Researchers must be aware of this disparity, and make efforts to ensure that ensuing results can still be validated with officially representative data.

The third chapter discussed the second category of development problems: Type II, or intermittent issues. While these kinds of problems might also be constantly occurring, measurements of them will have repeating, significant events over time, and it is the existence of these events that make the problem a development issue in the first place. Identifying early and measuring the extent of these events is information development practitioners need to help manage intermittent issues.

The three case studies chosen involve two dengue-based projects and one centered on the flu. Typically these disease-based issues are managed through disease surveillance systems, which involve a number of different on-the-ground reporting mechanisms. However, the complexity of these systems varies wildly country to country depending on level of infrastructure. Internet search terms and mobile phone networks are the discussed Big Data sources applied within the case studies to improve the dissemination of disease surveillance information. For the most part, the Big Data applications create improvements upon the traditional methods in terms of identifying and measuring the disease at hand. Speed is most important to those monitoring repetitive disease outbreaks. Identifying outbreaks early can be key to preventing injury or death, and understanding when and where the largest outbreaks are occurring as soon as they happen can help distribute health resources more efficiently. Therefore, improved speed is considered the primary benefit to Type II projects by Big Data. If the country in question in the case study had particularly lower infrastructure, the Big Data method also brought significant improvements to comprehensiveness and could even be considered an alternative to disease surveillance systems in areas where the creation of one is not a current possibility. However, this is less common.

Of all the available criticisms from Chapter 1, the issue of cultural relevance is likely most important here. Researchers often first calibrate their method at the beginning of projects by building an understanding of how intended populations interact with the technology being used where relevant. However, it is important to recognize that the culture around technology use will be constantly evolving, and that statistical models must be recalibrated over time to account for this change lest they lose accuracy. This is particularly challenging to do for intermittent issues. First, a loss in accuracy might not be noticeable until the next significant event occurs, potentially creating a hidden issue. Second, investigations into any cultural change in regards to use of related technology might only be possible in the middle of an event, offering less opportunities to make adjustments. Therefore, it is imperative researchers recognize the need in the first place to adjust their model along with social changes to maintain accuracy, and that they take every opportunity to do so.

Chapter 4 discussed Big Data responses to the third and final category of development problem, which are sudden crises or Type III issues. The main characteristic of these types of development issues are that they occur on relatively short notice. While they might repeat over time, even typically within a particular season, their occurrence is much less predictable than Type II issues or is not predictable at all.

Within the chapter, the three case studies offered deal with responses to an earthquake, floods, and a cholera infection outbreak using mobile phones, social media, and satellite signals. After a disaster strikes, response teams require data on the location of populations, risk, and damage. Traditionally this information is gathered through eyewitness accounts and by manually counting people who leave or enter aid areas, which can be slow, inaccurate, and reach only a small scope. In an emergency context when typically a large number of lives are immediately on the line, both the speed and comprehensiveness of this information is important. Applying various Big Data methods to these scenarios generally saw improvements to both. Populations could be tracked and risk assessed faster and sometimes at a wider scale.

Overall, however, Big Data applications to sudden crises had more mixed results than for Type I or Type II. One potential reason behind this is that for projects concerning ongoing or intermittent issues, researchers can for the most part choose a location that has an appropriate data set available. However, disasters will strike anywhere globally. Development practitioners must make do with what is available locally for sudden crises, which is not always best suited to help solve the particular issue at hand. Another issue is that disasters can actually remove access to typically available data sources. For example, mobile phone connectivity is sometimes reduced or eliminated following an earthquake. Some newer, more widespread, and more consistently available data sources like satellite measurements are starting to be tested, but the technology is not yet able to produce accurate measurements in all contexts. In total, there are more challenges in applying Big Data to Type III issues due to the needs of an emergency context and available technology that risk practitioners using a less appropriate dataset due to lack of options thereby reducing accuracy.

Implications, Limitations, and Recommendations

The general implication of these findings is that Big Data does have the potential to create significant, positive impact in the fields of development and humanitarian aid despite its many critiques. I argue that, as a result, Big Data for development efforts should be expanded. The field is extremely new, and thus it was difficult to assess actual impact of past work because few heavily studied applications have actually been put into practice. Many studies function simply as experiments to test if the use of a Big Data set can produce anything close to official development measurements. As demonstrated, the answer is usually positive. Future work in this sphere, built off of the promising foundation discussed in this thesis, is required to both build norms of operation and further trust in the method from funders and governments.

A few limitations to both the method of study and to the implications discussed exist. First, only academically researched forays into Big Data for development included enough detailed information to allow for assessment of impact within this thesis. Some for-profit and non-profit organizations are partaking in similar work, but it is not as heavily documented and the intentions behind the work can cloud investigations into its impact. Therefore, this cross section of Big Data for development work is important but was not covered.

It is also important to consider the limitations preventing further work from being accomplished in this arena regardless of potential for impact. All things considered, the engineering of solutions does not appear to be the most significant problem. Researchers in nearly every case took note of the most common criticisms and pitfalls associated with data for development work, took steps to avoid them where possible, and produced verifiably accurate results. A question for further research becomes, where is the disconnect? If Big Data for development has high potential for impact, and technology in many cases has proliferated enough through many developing country contexts to provide large datasets to use, why is it not utilized more?

A simple early answer is that the field is still growing. However, there are potentially larger, structural reasons that are less about growth or the technological strategies used and more about the people using them. If companies or institutions are unwilling to offer anonymized datasets to researchers, if governments, nonprofits, or universities lack the funding to support Big Data analytics, and if general mistrust and misuse of the technology continues, further utilization of Big Data in development efforts will be hindered.

However, there is an even more difficult question to answer. Even if all potential Big Data for development projects were supported, would governments and nonprofits have the capacity to actually use the results? In some cases the answer is yes, as discussed in the few case studies that had implemented results, but in others the answer is doubtful. The Big Data efforts discussed in this thesis are not taking place in a vacuum. There are larger issues generally within the field of development that if addressed might improve the capacity for impact of technologically-produced results.

Though, specifically to move forward progress on the Big Data front, a few adjustments and efforts could be made. Several common themes were found amongst case studies when making requests for future improvements in the field. First was a request to increase the number of partnerships between owners of large datasets and potential researchers. The more open the data is, the more opportunities exist to use it to improve people's lives, as long as privacy concerns are taken seriously and addressed. Second was continual improvement in the traditional methods used to solve development problems. Over time, practitioners are naturally getting better at measuring risk and delivering aid with the traditional methods, but the current system is not perfect, or even close. However, we must improve the traditional methods to improve the Big Data methods. Without valid data to test against, it is difficult to know if Big Data methods are an accurate alternative. The more accurate the traditional methods get, the more accurate Big

Data methods will be, as they are measured by the results produced from traditional methods. Finally, the most common future recommendation offered in the discussed case studies was a request for further research and validation based on promising preliminary results. Should these results continue to be confirmed into the future, Big Data will prove a valuable complement to existing international development and humanitarian aid efforts.

REFERENCES

Althouse BM, Ng YY, Cummings DAT (2011) Prediction of Dengue Incidence Using Search Query Surveillance. PLoS Negl Trop Dis 5(8): e1258. doi:10.1371/journal.pntd.0001258

Ashcroft, Michael, et al. "Detecting jihadist messages on twitter." *Intelligence and Security Informatics Conference (EISIC), 2015 European*. IEEE, 2015.

Bennett, James, and Stan Lanning. "The netflix prize." Proceedings of KDD cup and workshop. Vol. 2007. 2007.

Bizina, Margarita, and David H. Gray. "Radicalization of Youth as a Growing Concern for Counter-Terrorism Policy." *Global Security Studies* 5.1 (2014): 72-79.

Blumenstock, Joshua, and Lauren Fratamico. "Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data." *Proceedings of the 4th Annual Symposium on Computing for Development*. ACM, 2013.

Blumenstock, Joshua, and Nathan Eagle. "Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda." *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. ACM, 2010.

boyd, danah, and Kate Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, communication & society* 15.5 (2012): 662-679.

Buckee, Caroline O., et al. "Mobile Phones and Malaria: Modeling Human and Parasite Travel." *Travel Medicine and Infectious Disease*, vol. 11, no. 1,

2013., pp. 15-22
doi:http://dx.doi.org.proxy.lib.umich.edu/10.1016/j.tmaid.2012.12.003.

Burg, Steven L. "Preventing ethnic conflict: Macedonia and the pluralist paradigm." *Online: http://www. wilsoncenter. org/index. cfm* (1997).

"Cholera." World Health Organization. World Health Organization, Oct. 2016. Web. <http://www.who.int/mediacentre/factsheets/fs107/en/>.

Chunara, Rumi, Jason R. Andrews, and John S. Brownstein. "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak." *The American journal of tropical medicine and hygiene* 86.1 (2012): 39-45.

Cook, Samantha, et al. "Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic." *PloS one* 6.8 (2011): e23610.

"Communicable Disease Surveillance and Response Systems." *PHLS* (n.d.): 32. World Health Organization. Web. <http://www.who.int/csr/resources/publications/surveillance/WHO_CDS_EPR_LYO_2006_2.pdf>.

CRED, UNISDR. "The human cost of weather-related disasters, 1995–2015." *United Nations, Geneva* (2015).

Demchenko, Yuri, et al. "Addressing big data issues in scientific data infrastructure." *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013.

"Dengue and Severe Dengue." *World Health Organization*. World Health

Organization, n.d. Web.
<http://www.who.int/mediacentre/factsheets/fs117/en/>.

Doocy S, Daniels A, Murray S, Kirsch TD. The Human Impact of Floods: a Historical Review of Events 1980-2009 and Systematic Literature Review. PLOS Currents Disasters. 2013 Apr 16 . Edition 1. doi: 10.1371/currents.dis.f4deb457904936b07c09daa98ee8171a.

Ebach, Malte C., et al. "Big data and the historical sciences: A critique." *Geoforum* 71 (2016): 1-4.

"Fact Sheet about Malaria." *World Health Organization*. World Health Organization, n.d. Web.

"Flu Symptoms & Complications." *Centers for Disease Control and Prevention*. N.p., 23 May 2016. Web.
<https://www.cdc.gov/flu/about/disease/complications.htm>.

Hallegatte, Stephane, et al. *Unbreakable: Building the Resilience of the Poor in the Face of Natural Disasters*. World Bank Publications, 2016.

Lazer, David, et al. "The parable of Google flu: traps in big data analysis." *Science* 343.6176(2014): 1203-1205.

Letouze, Emmanuel. "Big Data for Development." UN Global Pulse, n.d. Web.
<http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf>

Narayanan, Arvind, and Vitaly Shmatikov. "How to break anonymity of the netflix prize dataset." *arXiv preprint cs/0610105* (2006).

Pallardy, Richard. "Haiti Earthquake of 2010." Encyclopædia Britannica. Encyclopædia Britannica, Inc., 18 Aug. 2016. Web.

<https://www.britannica.com/event/Haiti-earthquake-of-2010>.

Rehman, Nabeel Abdur, et al. "Fine-grained dengue forecasting using telephone triage services." *Science Advances* 2.7 (2016): e1501215.

Sakai, Yoko, et al. "Do Natural Disasters Affect the Poor Disproportionately? Price Change and Welfare Impact in the Aftermath of Typhoon Milenyo in the Rural Philippines." *World Development* (2017).

"Search Engine Market Share." NetMarketShare, n.d. Web.

Singel, Ryan. "Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims." *Wired*. Conde Nast, 17 Dec. 2009. Web. <https://www.wired.com/2009/12/netflix-privacy-lawsuit/>.

Smith, Tom W. "Measuring Racial and Ethnic Segregation." National Opinion Research Center, University of Chicago, Apr. 2002. Web. <http://gss.norc.org/Documents/reports/methodological-reports/MR096.pdf>.

Stevens, Philip. "Diseases of Poverty and the 10/90 Gap." *Diseases of Poverty and the 10/90 Gap* (n.d.): n. pag.*International Policy Network*. World Health Organization. Web.

"Tracking Flu Trends." *Official Google Blog*. N.p., 11 Nov. 2008. Web. <https://googleblog.blogspot.co.il/2008/11/tracking-flu-trends.html>.

"USAID's Office of U.S. Foreign Disaster Assistance Fact Sheet." *USAID* (n.d.): n. pag. 2017. Web. <https://www.usaid.gov/sites/default/files/documents/1866/ofda_fact_sheet_02-09-2017.pdf>.

Vigen, Tyler. "15 Insane Things That Correlate With Each Other." *Spurious Correlations*. N.p., n.d. Web.

<http://www.tylervigen.com/spurious-correlations>.

"What Climate Change Means for Africa, Asia and the Coastal Poor." World Bank. N.p., 19 June 2013. Web. <https://www.worldbank.org/en/news/feature/2013/06/19/what-climate-change-means-africa-asia-coastal-poor>.