



# BIG DATA

IN  
ACTION  
FOR  
DEVELOPMENT



**THE WORLD BANK**

IBRD • IDA | WORLD BANK GROUP

Latin America & the Caribbean

*Opportunities for All*



**WORLD BANK GROUP**

Macroeconomics & Fiscal Management

SECONDMUSE 

This volume is the result of a collaboration of World Bank staff (Andrea Coppola and Oscar Calvo-Gonzalez) and SecondMuse associates (Elizabeth Sabet, Natalia Arjomand, Ryan Siegel, Carrie Freeman, and Neisan Massarrat). The findings, interpretations, and conclusions expressed in this volume do not necessarily reflect the views of the Executive Directors of The World Bank or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work.

Design of the report by SecondMuse (Nick Skytland). The design of the “Data for Action Framework” and data palette used on case studies created by The Phuse.



# TABLE OF CONTENTS

Executive Summary	5
<b>Section 1. What is big data?</b>	8
The world is filled with data	8
Hello “big data”	9
Use and estimated value of big data	10
Big data in action for international development	11
<b>Section 2. How can we better understand and utilize big data?</b>	13
Insights and behaviors of interest	14
CASE STUDY: The Billion Prices Project and PriceStats	15
Generation to interpretation of data	18
Data generating process	18
Data content and structure	19
Data interpretation process	19
Insight implementation process	20
CASE STUDY: Understanding Labor Market Shocks using Mobile Phone Data	22
<b>Section 3. What can big data look like for the development sector?</b>	26
Examples, by medium and relevant data set	26
Big Data for Development in Central America: World Bank Pilot Efforts	27
Examples by medium and purpose	29
Areas of high potential for big data	30
Early warning	31
Enhancing awareness and enabling real-time feedback	31
Understanding and interacting with social systems	31
Specific challenges and ongoing processes	31
CASE STUDY: Forecasting and Awareness of Weather Patterns using Satellite Data	32
<b>Section 4. How can we work with big data?</b>	36
Technological capabilities	36
Human capabilities and data intermediaries	37
CASE STUDY: Connected Farmer Alliance	39
<b>Section 5. What are some of the challenges and considerations when working with big data?</b>	43
Data generation process and structure	43
Data interpretation process	44
Access	44
CASE STUDY: Tracking Food Price Inflation Using Twitter Data	46
Preparation	48
Analysis	48
Insights and their implementation	49
Insight scope and purpose	49
Insight implementation process	50
CASE STUDY: Using Google Trends to nowcast economic activity in Colombia	52
<b>Section 6. Retrospect and Prospect</b>	55
References	57
Annex 1: Selected Bibliography	60
Annex 2: Interview List	63
Annex 3: Glossary	65

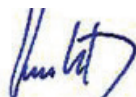
# FOREWORD

When we started this study, our main objective was to explore the potential of big data to close some of the existing data gaps in Central America. For us at the World Bank data are critical to design efficient and effective development policy recommendations, support their implementation, and evaluate results. Paradoxically many of the countries where poverty is high and hence good programs are more in need are also those countries where the data is more scarce. Not surprisingly then, we got seduced by the potential of big data. Not only has connectivity through access to mobile phones, internet, and social media increased dramatically; this has also generated an endless source of precious information. This has been well understood by marketing experts who now target individual needs. In addition to commercial use, these large amounts of data are a potential asset for the development community who could use them to help end poverty and promote shared prosperity. And indeed as this report documents, there are now good examples of how big data is used to improve growth forecasts by Ministries of Finance, track population movements, or plan emergency responses to weather related disasters.

In the end, we need to recognize that we did not advance as much as we wished in filling the blanks in our Central American Databases. True, we are working on three pilots that are helping us think about how to use these approaches in our daily work. For example, our teams are exploring whether we can use (i) night time illumination patterns captured by satellites to infer the spatial distribution of poverty; (ii) internet search keyword data to improve our forecasts of price series; and (iii) twitter data to better understand public reactions to policy decisions. Admittedly we did not have a major breakthrough but the work done helped us to start appreciating the potential of big data and we will continue pursuing this agenda trying to find country specific solutions that may emerge from big data analysis.

I consider worth sharing the work done by the team so far in (i) structuring a work program around a topic where we at the Bank had little, if any, expertise; (ii) presenting existing examples where big data is being used to improve development prospects, (iii) reflecting about the many development aspects that can be touched with this type of analysis, (iv) considering the technical and human capacity needs to make the best of big data, and (iv) assessing the challenges of working with big data. I also think there are important lessons emerging from the collaboration with SecondMuse and a good number of groups in the Bank: the staff in the Central American Country Management Unit, the Macroeconomics and Fiscal Management Global Practice, the Transport and ICT Global Practice, the Open Finances Group, the Development Economics Group, and the Innovation Lab in the World Bank Institute.

As the report notes, the age of big data is upon us. I hope that policy makers and development practitioners alike will find the work described in this report interesting and useful.



Humberto López

World Bank Country Director for Central America

# EXECUTIVE SUMMARY

This report stemmed from a World Bank pilot activity to explore the potential of big data to address development challenges in Central American countries. As part of this activity we collected and analyzed a number of examples of leveraging big data for development. Because of the growing interest in this topic this report makes available to a broader audience those examples as well as the underlying conceptual framework to think about big data for development.

To make effective use of big data, many practitioners emphasize the importance of beginning with a question instead of the data itself. A question clarifies the purpose of utilizing big data - whether it is for awareness, understanding, and/or forecasting. In addition, a question suggests the kinds of real-world behaviors or conditions that are of interest. These behaviors are encoded into data through some generating process which includes the media through which behavior is captured. Then various data sources are accessed, prepared, consolidated and analyzed. This ultimately gives rise to insights into the question of interest, which are implemented to effect changes in the relevant behaviors.

Big data is no panacea. Having a nuanced understanding of the challenges to apply big data in development will actually help to make the most of it. First, being able to count on plenty of data does not mean that you have the right data; and biased data could lead to misleading conclusions. Second, the risks of spurious correlations increase with the amount of data used. Third, sectoral expertise remains critical regardless of the amount of data available. And these are just some of the criticisms that one needs to take into account when thinking about the possibilities offered by big data. Care is needed in the use of big data and its interpretation, particularly since we are

still in an early stage for big data analytics.

In the development sector, various individuals and institutions are exploring the potential of big data. Call detail records via mobile phones are being used, often in combination with other data sources, to analyze population displacement, understand migration patterns, and improve emergency preparedness. Remote sensing images from satellites are showing promise to improve food security and minimize traffic congestion. Search queries and various text sources on the internet and social media are being ably analyzed for quick identification of disease epidemic changes or for inference of a population's sentiment on an event. Hence, big data shows promise to enhance real-time awareness, anticipate challenges, and deepen understanding of social systems by governments and other institutions. Yet, to move things forward will require the collaborative formulation of key questions of interest which lend themselves to the utilization of big data and the engagement of data scientists around the world to explore ways to address them.

Utilizing big data for any given endeavor requires a host of capabilities. Hardware and software capabilities are needed for interaction of data from a variety of sources in a way which is efficient and scalable. Human capabilities are needed not only to make sense of data but to ensure a question-centered approach, so that insights are actionable and relevant. To this end, cooperation between development experts as well as social scientists and computer scientists is extremely important [1].

Several challenges and considerations with big data must be kept in mind. This report touches on some of them and does not pretend to provide answers and solutions but rather to promote discussion.

For example, the process through which behaviors are encoded into data may have implications for the kinds of biases which must be accounted for when conducting statistical analyses. Data may be difficult to access, especially if it is held by private institutions. Even in the case of public institutions, datasets are often available but difficult to find due to limited metadata. Once data is opened, challenges around ensuring privacy and safety arise. This is also linked with the issue of personal data ownership. Even preparing data and ensuring its scalable and efficient use presents challenges such as the time and effort required to clean data. Analysis, especially when using big data to understand systems, must carefully consider modeling assumptions; algorithm transparency is critical to maximize trust in data driven intervention; and the work of translating insights into changes in the original behaviors of interest requires attention to those institutional structures and culture that will support the process.

The age of data is upon us. The means of its generation are undoubtedly multiplying, the technologies with which to analyze it are maturing, and efforts to apply such technologies to address social problems are emerging. Through a concerted and collaborative effort on the part of participants at various levels, big data can be utilized within meaningful systems and processes which seek to generate and apply insights to address the complex problems humanity faces.



## SECTION 1

# WHAT IS BIG DATA?

Data is a growing element of our lives. More and more data is being produced and becoming known in the popular literature as “big data”, its usage is becoming more pervasive, and its potential for international development is just beginning to be explored.

### The world is filled with data

In 2007, the world’s capacity to store data was just over  $10^{20}$  bytes, approaching the amount of information stored in a human’s DNA, and the numbers are growing. Between 1986 and 2007, the world’s capacity to store information increased by approximately four orders of magnitude. The technological capacity to process all this data has, in fact, experienced even more rapid growth [2]. To cite but one example of the growth rates, the Sloan Digital Sky Survey collected more data over a few weeks in the year 2000 than had been collected by astronomers to date [3].

Simultaneous with the rise of the capacity to store and analyze data is the increasing capacity for people to contribute to and access it. It is estimated that as of 2013, 39% of the world’s population has access to the internet. More astounding is the fact that mobile phone subscriptions are near 7 billion, approximately equaling the number of people on the face of the earth, yet access is not equal. While 75% of Europe’s population has access to the internet, only 16% have access in Africa. Interestingly, discrepancies for mobile phone subscription rates are not as pronounced. While Europe has approximately 126 subscriptions per 100 inhabitants, Africa has 63 [4]. Furthermore, the use of more advanced “smart” cell phones is

### KEY FINDINGS

- Big data can be used to enhance awareness (e.g. capturing population sentiments), understanding (e.g. explaining changes in food prices), and/or forecasting (e.g. predicting human migration patterns).
- Mediums that provide effective sources of big data include, inter alia, satellite, mobile phone, social media, internet text, internet search queries, and financial transactions. Added benefits accrue when data from various sources are carefully combined to create “mashups” which may reveal new insights.
- It is key to begin with questions, not with data. Once the setting for the analysis is defined, the focus of the research can move to the behaviors of interest and the consequent data generation process. The interpretation of this information will be used to produce actionable insights with the possible objective of influencing the behaviors of interest considered. Along these lines, this report develops a Data for Action Framework to better understand and utilize big data.
- Making good use of big data will require collaboration of various actors including data scientists and practitioners, leveraging their strengths to understand the technical possibilities as well as the context within which insights can be practically implemented.



expected to increase in the coming years. McKinsey Global Institute (McKinsey) estimates, for example, that for countries in Asia other than China, India or Japan, the number of “basic phones” is expected to decrease by 12% per year between 2010-2015, while “advanced phones” will rise by 17% during the same period [5].

## Hello “big data”

It is no surprise that, given such a rising tide, the horizon of possibilities considered by decision-makers over the years has increasingly taken into account how to make best use of such a deluge of data. As early as 1975, attendees at the Very Large Databases conference discussed how to manage the then-considered massive US census data [6]. In the late '90s, practitioners were already using massive, high-frequency store-level scanner data to compute optimal pricing schedules [7]. Indeed, it was around that time that the term “big data” was used to refer to the storage and analysis of large data collections [8]. Since the '90s, a plethora of media, often automatically capturing behaviors and conditions of people or places, provide new data sources for analysis. These media sources include online shopping websites capturing transaction data, retail computers capturing purchase data, internet-enabled devices capturing environmental data, mobile phones capturing location data, and social media capturing data on consumer sentiment.

Decreasing costs of storage and computing power have further stimulated the use of data-intensive decision making [8], and decision making based on ever larger and more complex datasets requires more sophisticated methods of analysis. For example, in the case of visual analysis, smaller datasets lend themselves to simple visual interpretations, say, using a scatter plot or line chart; while bigger datasets can't always readily be captured using similarly structured small and fast renderings [6].

As a result of the rapidly evolving landscape, the popular press, such as the New York Times [9], as well as academic discourses, have increasingly used the term “big data”, yet its definition has remained somewhat elusive. Technical audiences will often

distinguish big data with respect to the largeness of the dataset(s), say 200 gigabytes of data for a researcher in 2012. Practitioner audiences, on the other hand, will emphasize the value that comes from utilizing various kinds and sizes of datasets to make better decisions. Indeed, there does not appear to be any real and rigorous definition of big data; instead, it is often described in relative terms [3]. As an example, McKinsey uses the term big data to refer to datasets “whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”, thereby allowing the definition to vary by setting, such as industry, and time [5]. Yet with the great diversity of current storage and processing power, not to mention the doubling of such capacity in short time scales, such a definition makes comparability difficult --exactly the problem a definition seeks to avoid.

Several other authors [10], [11] often refer to the “Three V's” of big data: volume, variety, and velocity, originally discussed by Laney in 2001 [8], to distinguish big data. Volume refers to the actual size of the dataset(s) analyzed, variety to the various types of datasets possibly combined to produce new insights, and velocity to the frequency with which data is recorded and/or analyzed for action. The concept of variety often underlies use of the term big data. For example, in examining the use of big data for governance, Milakovich points out how “single sources of data are no longer sufficient to cope with the increasingly complicated problems in many policy arenas” [12]. In this vein, Boyd and Crawford point out that big data “is not notable because of its size, but because of its relationality to other data. Due to efforts to mine and aggregate data, Big Data is fundamentally networked” [13]. For the purposes of this report, which considers the development context, the use of big data will refer to the use of any dataset(s) which are distinguished by one or more of the three “V” features mentioned above for the purpose of generating actionable insights.

Other related terms such as open data and crowdsourced data have also become in vogue. “Open data”, for example, refers to data which is made technically and legally open, i.e. available in machine-readable format and licensed to permit

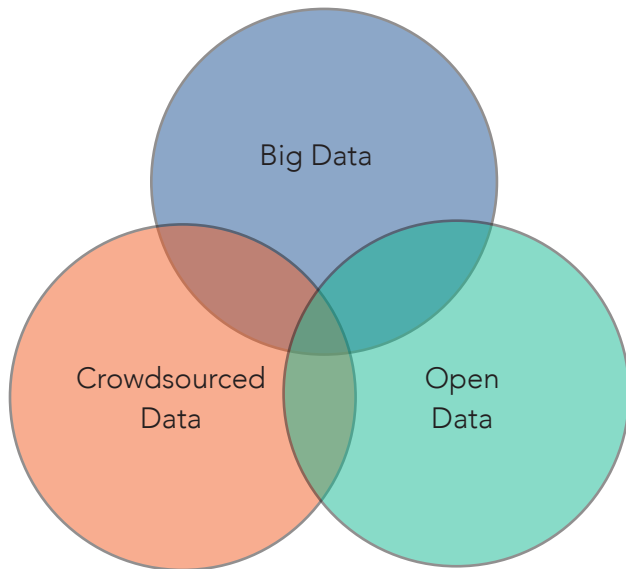


Figure 1: Relationship between Big Data, Crowdsourced Data, and Open Data

commercial and non-commercial utilization [14]. Cities such as New York sometimes open their data to stimulate innovation by drawing upon outside resources [15]. Countries such as Singapore, India, and the United States have also consolidated and opened data sets to the public. McKinsey distinguishes open data by explaining that, although big data may also be open data, it need not. Open data, they explain, refers to the degree to which data is liquid and transferrable [16]. Some data, particularly privately held mobile phone data for example, is particularly closed yet is certainly big data. Crowdsourced data is another popular term which refers to data collected through the aggregation of the input from large numbers of people. Crowdsourced data can also be big data, but need not be. “Crowdsourced” emphasizes the means through which data is collected whereas “big data” emphasizes the depth and complexity of the dataset(s).

## Use and estimated value of big data

The use of big data has, over the past several years, been motivated largely by private interests. In a survey done around 2010, with over 3000 business executives in over 100 countries, it was found that the top-performing organizations were “twice as likely to apply analytics to activities” including day-to-day operations and future strategies [17]. Businesses

have analyzed consumer purchase data to make personalized recommendations, video and shopping cart transponder data to streamline a grocery store’s layout, store- and product-level purchases together with climate data efficiently to maximize sales and minimize inventory costs, and location data on trucks together with geographic data to minimize fuel use and delivery times [5]. Accessing mobile data such as foot-traffic patterns or even phone operating systems have helped companies engage in more effective advertising [18]. Other specific examples include how Microsoft improved the accuracy of its grammar checker by increasing the relevant dataset from a million to a billion words, or how Google utilized a trillion words more effectively to provide language translation services [3]. In short, intelligent use of big data is becoming an effective way for companies to outperform their competitors, often through more effective foresight and understanding of the market dynamics [5]. While business has succeeded in demonstrating the marginal benefits which can accrue from big data, e.g. efficiency gains in more effective retail forecasting, more substantive or qualitative effects of big data, particularly in terms of social practices, are just beginning to emerge [19]. That said, the marginal benefits are substantial for private and public interests.

McKinsey estimates the potential additional value of big data in the sectors of US health care, European public sector administration, global personal location data, US retail, and global manufacturing to be over \$1 trillion US dollars per year, half of which comes from manufacturing alone. Such value often comes from efficiency gains by reducing the inputs required for the same amount of output [5]. Another study estimated the value of big data via improvements in customer intelligence, supply chain intelligence, performance improvements, fraud detection, as well as quality and risk management. For the United Kingdom alone, this value was estimated at \$41 billion US dollars per year in the private and public sectors [20].

## Big data in action for international development

In addition to providing insight to make businesses more profitable, big data is showing promise to improve, and perhaps substantively change, the international development sector in novel ways [10]. Of general interest is the fact that big data often is produced at a much more disaggregated level, e.g. individual instead of, say, a country level. Whereas aggregated data glosses over the often wide-ranging disparities within a population, disaggregated data allows decision makers more objectively to consider those portions of the population who were previously neglected [21].

Two basic approaches appear to stand out with respect to big data in an international development context. One is when big data is utilized for projects or processes which seek to analyze behaviors outside of government or development agencies in order to heighten awareness and inform decision making. Another approach is when big data is utilized for the analysis of behaviors internal to a single institution, such as the government, often to streamline and improve services.

Several examples demonstrate the use of big data for projects or processes more concerned with matters outside of governments or other types of agencies. For example, in one case it was found that a country's gross domestic product could be estimated in real-time using light emission data collected via remote sensing. In this way, alternate data sources serve as a proxy for official statistics. This is especially helpful in the development context since there is often a scarcity of reliable quantitative data in such settings [22]. Changes in the number of Tweets mentioning the price of rice in Indonesia were closely correlated to more directly measured indicators for food price inflation. Similarly, conversations about employment including the sentiment of "confusion" via blogs, online forums, and news conversations in Ireland were found to precede by three months official statistics showing increases in unemployment [10]. In Guatemala, a pilot project explored how mobile phone movement patterns could be used to predict socioeconomic status, thereby approximating census

maps which are often prohibitively expensive to produce. In Mexico, analysis of call detail records enabled tracking of population movements in response to the spread of epidemic disease and provided insight into the impact of policy levers like transportation hub closures, such that the velocity of infection rates was reduced by as much as 40 hours [23]. In Kenya, once the impact of mobile money transfers was evident, governmental regulations were changed to enable their increased use [18]. Text analysis of social media data has the potential to identify issues pertaining to various population segments over time, e.g. refugee challenges or political opinions, thereby allowing development organizations more effectively to listen to the needs of a population [24].

Several governments have used big data in a variety of ways to streamline processes, thereby increasing cost and time savings. In Sweden, the government used previous years' data combined with user confirmation via text messaging to streamline tax filings. In Germany, the Federal Labor Agency used its multidimensional historical customer data more effectively to assist unemployed workers, thereby reducing costs by approximately \$15 billion USD annually. Altogether, McKinsey estimates that Europe's 23 largest governments could create \$200-400 billion USD per year in new value over 10 years through the use of big data to reduce mistakes and fraudulent tax reporting behaviors [5].

The examples above may be used to stimulate thinking on similar or analogous uses of big data to drive resource efficiency, process innovation, and citizen involvement where resources are constrained, thus laying a strong foundation for poverty alleviation and shared prosperity around the world.



281-4  
291-4  
281-4  
305-4  
319-0  
329 0

28  
25  
26  
27  
28  
29

## SECTION 2

# HOW CAN WE BETTER UNDERSTAND AND UTILIZE BIG DATA?

Before delving into a technical and political analysis of the use of big data, it is helpful to have a contextual understanding of how such data may be generated, accessed and acted upon. Figure 2 below describes this framework of data for action. This follows from the recognition that data is not the same as knowledge and that, therefore, a whole host of capacities are required to generate data-driven actionable insights for social betterment.

At the most basic level, there are behaviors or conditions existing in the world which include, among many other things, climate, human sentiments, population movement, demographics, infrastructure and market-based interactions. These behaviors and conditions, named behaviors henceforth, are encoded through some data generating process which includes triggers and media through which data is recorded. Once data in whatever form is generated,

a data interpretation process takes place through which the raw data is accessed, consolidated, and analyzed to produce some actionable insight. Often, this interpretation process is cyclical and interactive rather than strictly linear. For example, the analysis may shed light on new data needed, thus requiring access to new data, or the very act of consolidating data may already reveal insights which will inform the analysis or the need to access data in a different way. Once actionable insights are determined, they must change the behaviors of interest through some implementation process. This process includes distilling insights into next steps, structuring organizational processes or projects accordingly, and engaging in the corresponding actions. To the extent that the insights gained are accurate and the implementation is done thoroughly, the behaviors will change as expected. In whatever case, however, all phases of the cycle inform each other through a rich process of learning.

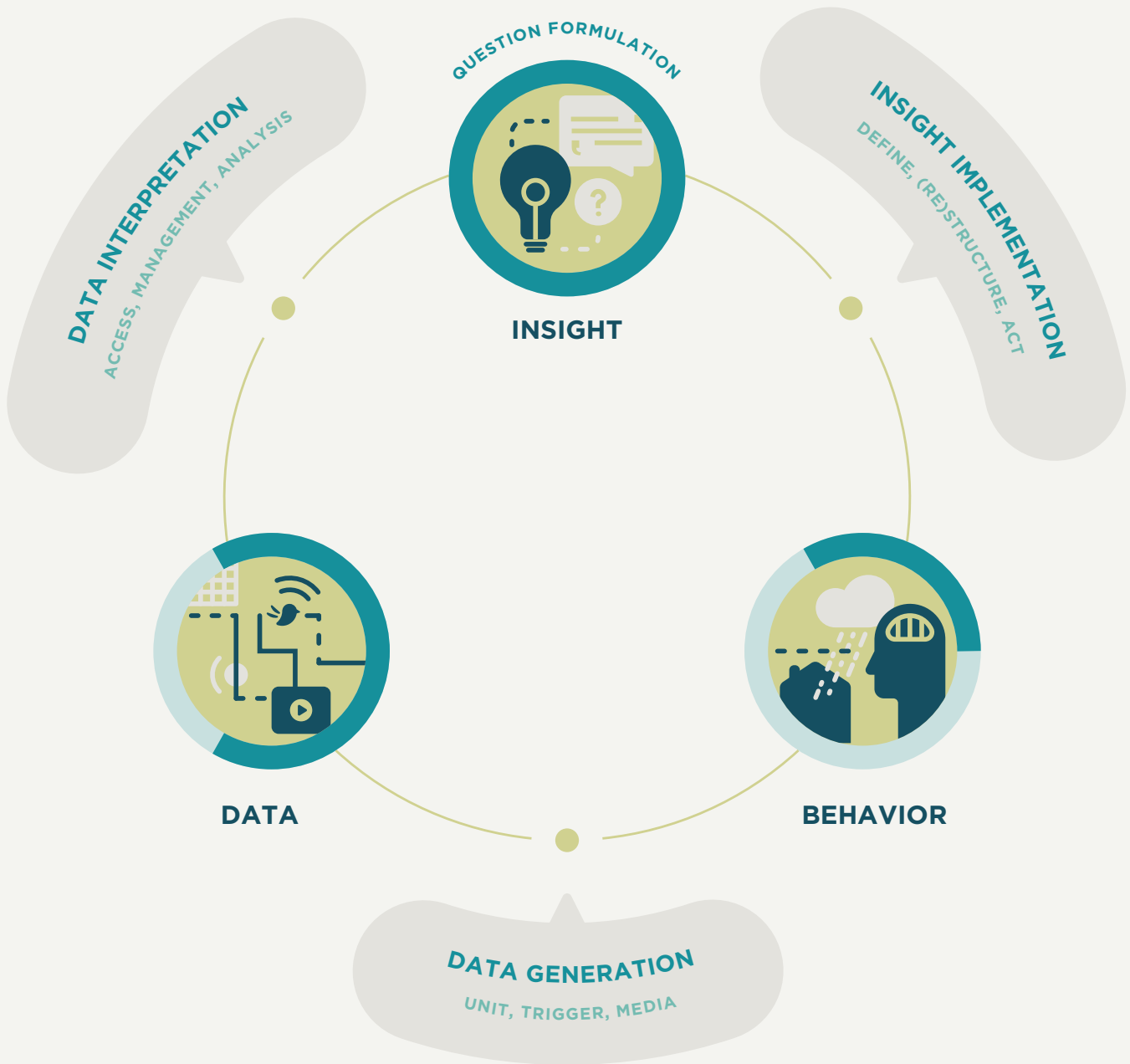


Figure 2: Data for Action Framework

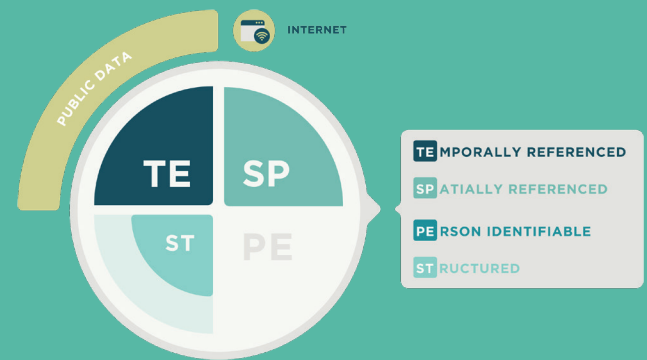
The above framework is general enough to permit an exploration of the use of big data in a variety of contexts yet also specific enough to frame the subsequent discussion. Each element of the figure above is discussed in further detail in the sections below, starting with a careful consideration for the “end” use of such data, i.e. the kinds of insights and actions we wish to generate.

## Insights and behaviors of interest

As expressed by experts who were interviewed for this report [25], [26], [27], [28], [29] as well as in a report distilling experience for over 3,000 business executives around the world [17], the best way to proceed when considering how to use big data is to begin with questions, not data. Undoubtedly, as data is collected and analyzed, a question will be refined.

# The Billion Prices Project and PriceStats

Retail prices via websites



## Motivation

The Billion Prices Project began when Alberto Cavallo, Sloan Faculty Director at MIT, noticed that Argentina could benefit from a more transparent, reliable, and low-cost method to track inflation. The primary purpose, therefore, was to enhance awareness on price changes and purchase behaviors of interest, particularly those prices consumers were paying. A methodological challenge was also identified: could data via the Internet feasibly and reliably provide an alternative data source for traditional price index measures? In time, as the data and methods to access them proved reliable, the data could be used by researchers to enhance understanding as well as central banks and financial traders to enhance their forecasting abilities.

Online price data provide distinct advantages over alternative sources of information such as direct survey methods used for traditional price indices or scanner data obtained through companies. In particular, online price data is very high-frequency (daily)--in fact is available in real-time without delay--and has detailed product information such as the size or exact product identifier (i.e. SKU) as well as whether the product is on sale. Unfortunately, data on quantity sold is relatively unavailable via outward-facing Internet webpages.

## Data Generation

As indicated by the data palette above, the data used for this study is generated by various firms via the Internet. While procedures to update prices will vary across companies it is safe to state that, in general, such information is provided on a high-frequency and specific geographical basis (i.e. with store location), or as indicated by the palette, at a high level of temporal frequency and spatial granularity. The data is relatively structured since contextual text on the webpage will indicate the SKU for the product as well as other associated parameters. Once posted online, such data become publically accessible for anyone to access.

## Data Interpretation

In the initial efforts to compute price indices for the country of Argentina, all Dr. Cavallo needed was his laptop and time. As the endeavor grew in scope, however, methods and considerations grew in sophistication. Today, a curation process ensures that the best sources of online data are selected. In some cases, data is collected from a retailer for a few years to evaluate whether the quality of the data is sufficiently high for inclusion. Also important is to capture offline retailers' price information since fewer individuals in developing countries, in particular, purchase goods online. Although each download or "scraping" places no more pressure on a server than a regular page view, the timing is even adjusted to reduce retailers' server demands. Finally, retailers'

privacy concerns are addressed in a variety of ways including sharing data at an aggregated level and with time lags.

While the technology for “scraping” price and associated product information off of websites is inexpensive and readily available, Dr. Cavallo and his team realized that a careful process of cleaning the pulled data needed to be put in place; this to ensure that all data sources are homogenized and prepared for aggregation and analysis.

The consolidated price data has been used in two ways. The first is to produce price indices which give an indication of inflation. This information is sold and/or shared with central banks and financial traders. The second is for academics seeking to understand market behaviors. Academics will often utilize econometric techniques, for example, which leverage the data sets’ high degree of granularity and dimensionality in order to understand short run or disaggregated effects of various policies.

## Insight Implementation

Individuals working at central banks and/or traders use such data to enhance decision making. For example, due to its high-frequency, central banks can see day-to-day volatility combined with sector-by-sector comparisons which traditional measures can’t capture. As experience with such data sources deepens, many governmental statistical offices are shifting their mentality to accept alternative methods of data collection.

Academics also use such data to conduct research. The fact that the data has so many dimensions allows economists to avoid the use of complicated techniques to account for expected gaps of information in traditional data sources. In one paper, economists used the Internet scraped data to show that the Law of One Price (a central economic theoretical result) tends to hold across countries that belong to formal or informal currency unions. In another paper, the dataset was used to show how natural disasters affect the availability and prices of goods. Researchers found that more indispensable goods were prone to more quickly reduce in availability and more slowly to resume from price increases.

## From Idea to Ongoing Process

What started as an idea by one individual grew, in time, to a research initiative named Billion Prices Project at MIT which was primarily supported by grants. In time, the company PriceStats was created through which high-frequency and industry-specific price indices could be sold. Through an agreement between PriceStats and the Billion Price Project to share the data as well as some of the earnings, the research work can continue uninterrupted. In time, it is anticipated that the data will be more widely shared with academic institutions for research purposes.

## References

- BPP. (2014). The Billion Prices Project @ MIT. Retrieved from <http://bpp.mit.edu/usa/>
- Cavallo, A., Neiman, B., & Rigobon, R. (2012). Currency Unions, Product Introductions, and the Real Exchange Rate (No. w18563). National Bureau of Economic Research
- Cavallo, A., Cavallo, E., & Rigobon, R. (2013). Prices and Supply Disruptions during Natural Disasters (No. w19474). National Bureau of Economic Research.
- Conversation with Alberto Cavallo, February 2014
- PriceStats. (2014). History. Retrieved from <http://www.pricestats.com/about-us/history>



A well defined question will clarify three basic elements: the primary purpose for using big data, the kinds of real-world behaviors of interest, and the scope of analysis. The following are possible questions. How do incomes in Uganda change throughout the year? How can assistance be more effectively provided to those in need after a natural disaster near the Pacific Ocean? What cities in the world should prepare for increased flooding due to climate change? How can governmental systems be designed to efficiently and appropriately tax individuals according to transparently defined principles? How can a country's employment sector be more effectively coordinated with training and education?

An analysis of multiple case studies, such as the ones included in the introduction, suggests that endeavors which utilize big data are primarily focused on advancing either awareness, understanding, or forecasting. Use of big data to heighten awareness is exemplified by projects which utilize non traditional sources of data to serve as proxies for official statistics, such as the gross domestic product or Kenyan examples above. Real-time monitoring of disasters provides another avenue in which awareness is especially needed [30]. Big data in some cases is used to more deeply understand a phenomenon so that better policy levers can be utilized. The Mexico call detail records case described above is one relevant example. Finally, big data may be utilized to more accurately forecast behaviors so that institutions and populations can more effectively prepare. The unemployment case in Ireland is one such example. Without a doubt, these three purposes are deeply interrelated. It is impossible to advance understanding, for example, without heightening awareness. Understanding, in turn, is often the foundation upon which forecasting methods are utilized. Conversely, certain machine-learning and inductive algorithms may be used to enhance forecasting ability which can itself give rise to a better understanding of a system. That said, the three categories may be useful to stimulate thinking as far as the ways in which big data can be utilized. Awareness, understanding, and forecasting aptly

correspond, for example, to the actions institutions may take to respond to present situations, design policy, or prepare for future events.



Figure 3: Awareness/Understanding/Forecasting

Regarding the question of behaviors, the following categories (with corresponding examples) are suggestive of areas of interest in the international socio-economic development context. The table below shows how a few of the example behaviors may be used in the context of heightening awareness, advancing understanding, or enhancing forecasting.

- product/service usage (e.g. non-market food consumption)
- market transactions (e.g. wheat purchase prices)
- human population movement (e.g. regional migration patterns)
- human population sentiment (e.g. public opinion on policies)
- human population conditions (e.g. extent of disease epidemic)
- weather conditions (e.g. ground temperatures)
- natural resource conditions (e.g. extent of forests)
- physical infrastructure conditions (e.g. locations of usable roads)
- agricultural production (e.g. extent of rice cultivation)

	AWARENESS	UNDERSTANDING	FORECASTING
<b>Wheat purchase prices</b>	How much are farmers currently receiving for the wheat they are selling?	What is driving changes in wheat purchase prices?	What will wheat purchase prices be in a week?
<b>Public opinion on policies</b>	How favorably do citizens feel about a particular policy?	What factors drive public opinion on foreign relation policies?	How will public opinion change in the coming months?
<b>Regional migration patterns</b>	During what times of the year do people engage in migration	How do labor wage differences stimulate changes in migration patterns?	How is country-to-country migration expected to change in the next few years?

In determining the scope of analysis, it is helpful to know if the use of big data in a process or project is intended to consider the situation for a single individual, city, region, nation, and/or the entire planet. By determining the scope, data requirements as far as size and interoperability become clear.

## Generation to interpretation of data

Once the setting for the analysis is well defined, it will be helpful to consider the data available corresponding to the behaviors of interest. The first step in this is to describe and categorize the various types of data available.

Many authors have attempted to do this. One report notes that data may record what people say or do [32]. Another report points out that big data may have one or more of several features including whether it is digitally generated, passively produced, automatically collected, or geographically or temporally trackable, and/or continuously analyzed. The same report discusses a data taxonomy with four categories: data exhaust (i.e. passively generated, often real-time data), online information, physical sensors, and citizen reported or crowdsourced data. Also included in the report is the division of data sources into traditional (e.g. census or survey data) vs. nontraditional (e.g. social media, mobile phone data) [10]. Another paper discusses how data may be: raw (primary, unprocessed data directly from the source), real-time data (measured and accessible with minimal delay),

and/or linked data (published in a format which lends itself to identify elements and links between datasets) [33]. Yet another paper explores how data may be structured (i.e. readily stored and accessed in terms of columns and rows) or unstructured (e.g. images or video) [34].

An examination of the above categories reveals that underlying the discourse to understand big data are three aspects of the cycle described above: the way in which data is generated, its content and structure, and the process through which it is accessed and interpreted. The following sections elucidate these three aspects, providing relevant categories with which to organize thinking about the opportunities in the big data space.

### *Data generating process*

The data generating process has at least three features: the data-recording trigger, the level at which data is collected, and media through which data is generated. To begin, the trigger that is utilized to encode behaviors into data may be active or passive. For example, much mobile data is passively or constantly collected by cell phone towers. Data such as a Twitter or Facebook post is generated actively since a human decision actively precipitates its recording. Data is also generated at a particular, most granular level of analysis. These levels of analysis may pertain to temporal, spatial, human, or other characteristics. For example, retail sales data

is generated at a high temporal frequency (daily, in some cases), spatially identified at a store-level (i.e. latitude and longitude), and by product SKU. Finally, the media through which data is generated may include one or more of the following: satellite, mobile phone, point-of-sale, internet purchase, environmental sensors, and social media, among others.

## Data content and structure

In terms of the data content and structure itself, at least six features shed light on the kind of big data one may choose. First, as mentioned above, data may be in a structured or unstructured form. Second, data may be temporally-referenced; in other words, each record or instance has some form of temporal identification attached. Third, data may be spatially-referenced, i.e. be tied to some geographic location data. Fourth, data may be person-identifiable; in other words, records are not only separable<sup>1</sup> and unique by person, but fail the test of anonymity. Fifth, data may have various sizes, from, say, hundreds of megabytes to several petabytes. Finally, a dataset may or may not be a compilation of other datasets.

## Data interpretation process

Once datasets corresponding to the behaviors of interest have been identified, these must be collected and analyzed so that actionable insights may be generated.

Several authors have elucidated features pertaining specifically to the process of analyzing data. In one paper, several phases are described, including: data acquisition, information extraction and cleaning; integration, aggregation, and representation of the information; query processing, data modeling, analysis; and, finally, interpretation. Each of these phases presents challenges such as the heterogeneity of the data sources, the scale of the data, the speed at which data is generated and response is needed, concerns for privacy, as well as enabling

human collaboration [35]. A study done with sixteen data analysts using big data at Microsoft found five major steps which most engaged in: acquiring data, choosing an architecture (based on cost and performance), shaping the data to the architecture, writing and editing code, and finally reflecting and iterating on the results [6]. Finally, another report by

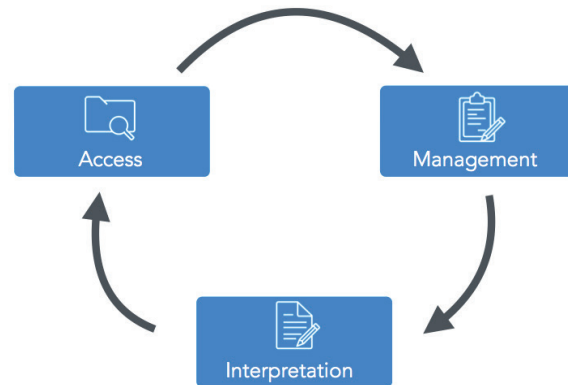


Figure 4: Data Interpretation Process

the United Nations discusses how big data analysis requires the steps of filtering (i.e. keeping only relevant observations), summarizing, and categorizing the data [10].

In reviewing the above and several other papers, three interrelated components are involved in generating insights from data. First, the data must be accessed. Second, the data must be prepared. Third, the data must be analyzed. Undoubtedly, these three components are highly interrelated and need not progress linearly. For example, analysis can inform the need to access new data, and the data preparation component may reveal certain preliminary insights for analysis.

### Access

In categorizing the ways to work with big data, many sources have actually been describing features pertaining to the access and/or generation of data. At a most basic level, data may be accessed using one of three sources: via public institutions, via private institutions, and directly from individuals or the crowd [36]. Publicly-sourced data includes zip-code level US census data, or massive weather station data from

1. Separability implies that an individual's records can be separated and distinguished from others, whereas identifiability implies that it is known who that individual--often simply represented by a number of code in a database--is. As long as a database is person-identifiable it is individually separable. However a database could include separable yet not identifiable individuals. One such example is an individual wage record database where individuals' names have been scrambled with random, yet consistent identifiers. It is worth noting, however, that even in cases where such data has been scrambled, there are settings and corresponding methods through which an analyst could inductively identify people.

the US National Oceanic Atmospheric Administration (NOAA). Privately-sourced data may include store-level retail sales data for hundreds of thousands of stores across a country or mobile phone location data for millions of individuals across a country. Crowd-sourced data may include millions of image artifact-identification analyses done by hundreds of thousands of individuals.

### **Management**

Given that one or more dataset(s) can be created and/or accessed using the means discussed above, the data needs to be combined and prepared for analysis. At this stage, the steps mentioned above such as choosing an architecture, extracting and cleaning the data, and integrating the datasets become relevant. This component of process, which will be further discussed in Section 4, takes a great deal of time and is closely tied with interpretation.

### **Interpretation**

Once data has been formatted in such a way that it can be more readily accessed by the analyst, various methods can be used to interact with it and generate insights, or at least elucidate other questions which will refine data access and preparation. Two main mechanisms are often used to interpret big data: visualization and modeling.

Visualization plays a key role since it leverages the human strength to see patterns. It also often helps the analyst scrutinize the data more closely [37] and may enable ready comprehension on findings that would otherwise be difficult to achieve [10]. Indeed, in a survey of over 3000 business executives around the world, a great number indicated that “visualizing data differently will become increasingly valuable” [17]. Visualization requires the thoughtful selection of the relevant pieces of information displayed in a visually appealing way to help a decision-maker understand the data.

Modeling is essential to interpreting data, especially if the purpose is to understand and/or forecast behaviors. Models attempt to describe the underlying processes occurring in the world which give rise to the

behaviors of interest. Through the lens of a particular model, the data sheds light on the presence and degree to which relationships exist. For this reason, model selection is especially important, otherwise true relationships may remain undetected. To ensure a proper model, many researchers will emphasize how experience, expertise and human intuition are critical [38]. In addition, it is important to consider the fact that, when modeling human, non-laboratory/controlled settings using high-frequency big data, several constraints and parameters affect the behaviors of interest. In this regard, models must be thoughtfully designed [39]. From a statistical-scientific standpoint, the use of big data has significant implications for modeling and theory-development. At a basic level, an analysis of substantial amounts of data can inform the design of models and vice versa such that a dynamic interplay exists between them [37].

### **Insight implementation process**

Simply having insights pertaining to relevant behaviors is insufficient to cause a change in those behaviors. A process whereby insights generated are translated into action is necessary. In considering this process, at least three features stand out: defining next steps, adjusting or creating structures to ensure these steps are carried out, and taking the necessary actions.

The insights which are ultimately generated as a result of the analysis of data may be more or less actionable. Therefore, it is critical to articulate next steps arising from the insights. And, of course, to act upon these steps. However, it is also necessary that some kind of structure is in place to ensure continuity of action.

In this regard, the distinction between using big data for one-time projects versus integrating its usage into an ongoing process merits consideration. Many short-term projects have utilized big data to explore its application as a proxy for official statistics or, in some cases, as a way to allocate resources after a disaster. Once immediate disaster relief is carried out, for example, continued use and refinement of the original big data may cease. At best the data is used as reference material for future projects.

However, big data may be used in the context of ongoing processes through which its use is refined and the insights it leads to are ongoingly acted upon. Retail giants like Walmart provide examples of the ways in which big data can be integrated in the ongoing functioning of the company such as using forecasts based upon their large stores of data in order to adjust inventories, re-organize content, and price items. Cities like Chicago are gathering and tracking both historical and real-time big data on an ongoing basis to streamline operations and uncover meaningful correlations [31].

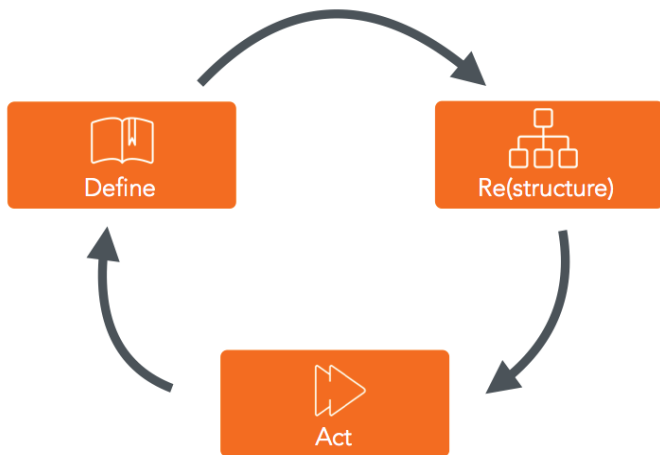


Figure 5: Insight Implementation Process

Whatever the nature of the use of big data, whether for a project or process, the three facets of the implementation process--defining, (re)structuring, and

acting--are closely tied and are not necessary linear. It's possible that action on the ground informs how structure is defined. For example, when aid agencies assist potential polio victims, they may discover that methods of communication need enhancing in particular ways. Alternatively, creating a structure to implement next steps may itself help lend further shape to them. For example, a non-profit may see that the insights generated from an analysis very clearly point to the need to increase the number of vaccinations. However, when beginning to define the structures which will actually carry this out, it may be discovered that other organizations are already doing this and that what is actually necessary is to more appropriately identify those individuals who need vaccinations. Critical to ensuring that the process has a healthy degree of interactivity among its various elements is a culture of learning characterized by a willingness to share observations and learn from mistakes.

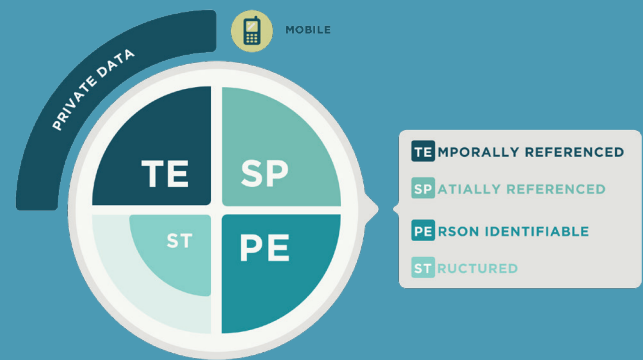
Once insights are generated and implemented, behaviors change and new data is generated, whereby the cycle resumes and a rich process of learning ensues.

# CASE STUDY

## DATA PALETTES

# Understanding Labor Market Shocks using Mobile Phone Data

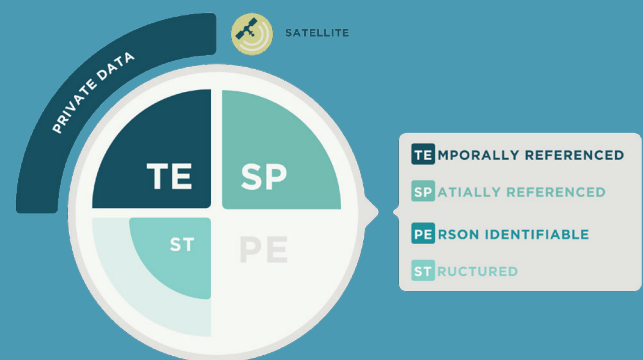
Call detail records via mobile phones



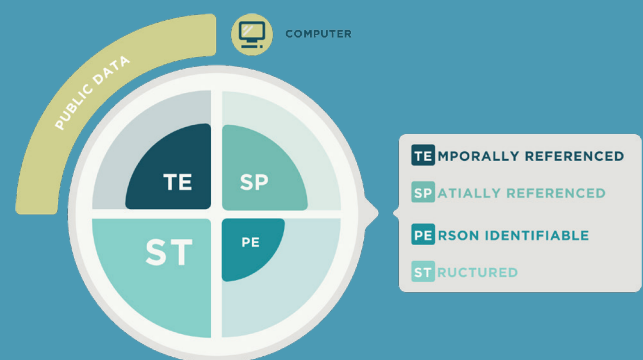
## Motivation

The use of call detail records (CDRs) to track fine grain patterns of population movement has been a topic of much research in the past few years. Joshua Blumenstock, University of Washington professor, has spearheaded projects in Rwanda and Afghanistan using CDR records which enhance the ability to gather deeper insights into more nuanced forms of migration, including seasonal and temporary migration. Upon the basis of that research, Blumenstock and his colleague Dave Donaldson, are now pursuing evidence to support the long-held theory of migrant workers acting as arbitrageurs whose movement among labor markets serve to bring those markets into equilibrium. The highly-detailed internal migration data necessary to support that theory empirically has been unavailable until now. In addressing their central research question regarding the extent to which the movement of people over space helps stabilize and equilibrate wage dispersion, Blumenstock and Donaldson are seeking to analyze migration patterns in response to labor market shocks to shed light on the dynamics of internal migration in low income economies.

Weather conditions via remote sensing imagery



Labor market data via public agency surveys



## Data Generation

Although access to CDR data is typically a significant hurdle in projects of this kind, in this case Blumenstock and Donaldson already had access to the necessary CDR data from prior projects. Data is generated automatically as telecommunications companies encode phone calls into mobile phone records. For this study, the data spans several years, accounts for three developing economies and is highly detailed both

temporally and spatially. In addition to CDR data, the project is making use of census data and external data sources underlying labor demand shocks, including weather conditions and domestic and international commodity prices. Public census data, while it changes rarely, includes a high volume of information and adds depth of knowledge to the longer term dynamics of rural to urban area migration. Government labor market data is highly structured and includes regional price and wage information at two regular intervals. Finally, the team purchased the high frequency/high resolution satellite weather data needed to assess climate-related labor shocks for the project.

## Data Interpretation

CDR data is a remarkably unwieldy and inconsistent dataset to work with, received in unstructured repositories of millions of individual files. The initial steps in the project from pre-processing the data to teasing out the relationships between the different datasets (e.g. wage and crop price data) are incredibly time consuming. The team works with the data on a daily basis and builds models iteratively based on the data.

The data analysis process goes through three steps. The first step is to isolate local shocks to labor demand (e.g. weather or commodity prices changes) and to identify the resulting labor market outcomes. The second step involves using the identified shocks to labor demand and drawing on CDR data to estimate the migration response to those shocks. This step aims to understand the migration response to labor market shocks and answer the following questions: What types of individuals are more or less likely to migrate in response to shocks? What regions are more receptive to migrants? How long after a shock do people stop migrating? How do these dynamics affect urbanization? The final step in the analysis is to estimate the actual effects of migration dynamics on the creation of wage equilibrium and to understand the speed and depth of the impact.

There is already significant existing research on models used to determine measures of mobility using CDR data and several research papers on the topic. Moreover, the project draws on classic theoretical models from economic literature, like the Harris-Todaro model, regarding the relationship between wages, labor market shocks and migration. The project team is currently building a quantitative framework that will allow them to test those theories based on the iterative models being developed through analysis of the wage and migration data. Throughout the process of analysis, small insights continuously hone the questions asked and help determine relevant inputs. As data comes in, determinations are continuously made which refine the quantitative models that exist for each factor within the framework and the relationship between factors. Benchmarking against other existing data sources, including census data, is relevant in making those determinations.

## Insight Implementation

Though this project is still in its early stages, it will deepen understanding of the role of migrants in labor markets. While the theories underlying many current policies designed to impact migration and urban-rural population flow are well-developed, empirical evidence is remarkably thin due to the lack of detailed data. As tracking of population movements using CDR data becomes more commonplace, the insights gained through this historical analysis of data will play an integral part in understanding migration patterns.

## From Idea to Ongoing Process

Like many of Blumenstock's projects which track migration patterns on the basis of CDR data, this retrospective project contributes to a fundamentally deeper understanding of how wages and labor markets are determined. Through such an understanding, labor policy can be more effectively designed for low income countries, including selection of actions to incentivize and disincentivize migratory behavior. By pioneering, and rigorously documenting, a process for gathering insights based on a quantitative framework of evidence, this project could be a foundation upon which ongoing evaluation of government policies could be conducted.

## References

Blumenstock, J. & Donaldson, D., (2013). How Do Labor Markets Equilibrate? Using Mobile Phone Records to Estimate the Effect of Local Labor Demand Shocks on Internal Migration and Local Wages. Proposal Summary for Application C2-RA4-205.

Conversation with Joshua Blumenstock, March 2014.

Harris, J. & Todaro, M., (1970). Migration, Unemployment and Development: A Two Sector Analysis. *American Economic Review* 60 (1).





ro. Tareem Ph.: 9873484265

**AASHO FABRICATION & WELDING WORKS**

Mfg. OF :-  
 GATE GRILL, WINDOW JALI  
 A.C. STAND, COOLER STAND

Spl in :-  
 SIDE FABRICATION WORKS  
 TIN SHADS & FIBER SHEET SHADS

No. 20, LN.J.P. MARKET RANJEET SINGH MARKET, NEW DELHI-2

**STD  
 ISD  
 PCO**  
 AASHO COMMUNICATION

**STD  
 ISD  
 PCO**

**STD  
 ISD  
 PCO**

**STD  
 ISD  
 PCO**

**Airtel**

## SECTION 3

# WHAT CAN BIG DATA LOOK LIKE FOR THE DEVELOPMENT SECTOR?

Big data shows potential to advance development work in a variety of ways. In the first section above, several examples were provided which highlighted the ways in which big data could be used as a proxy for conventional official statistics, thereby enhancing institutional awareness of the conditions of a population; to better organize governmental processes thereby delivering more effective services; or to enhance understanding of the drivers of health epidemics, thereby guiding policy decisions.

Any point in the framework discussed in Section 2 can be used to stimulate the imagination on the horizon of possibilities for big data in development. The case studies presented throughout the text provide concrete applications of the lens of the framework in various settings. Moreover, this section describes several examples of data sets utilized by medium as well as by purpose and report information on the first World Bank attempts to leverage big data to address development challenges in Central American countries. By cross referencing primary media with the primary purpose of the use of big data--awareness, understanding, or forecasting--one can easily see how big data projects can take a variety of configurations depending on the context. Then a summary is presented detailing what institutions and individuals are saying about where big data shows promise for development. Finally, recommendations for next steps in advancing the application of big data for development are provided.

### Examples by medium and relevant data set

**Mobile | Call Detail Records.** Although usage of call detail record (CDR) data for development is still in early phases, applications such as using Digicel's data to track population displacement after the Haiti earthquake and modeling of infectious disease spread show great promise [40]. One study in Afghanistan showed that CDR data could be used to detect impacts from small-scale violence, such as skirmishes and improvised explosive devices, in terms of their impacts on communication methods and patterns of mobility. Another project done by the lead researcher in the Afghanistan study was to capture seasonal and temporary migration, usually overlooked by traditional survey models, permitting a more precise quantification of its prevalence. An ongoing project which builds upon these results aims to measure precisely the extent to which wage disparities in Rwanda, Afghanistan, and Pakistan are arbitrated by migration [22].

**Satellite | remote sensing images.** Usage of satellite data abounds. For example, the United Nations University engaged in a project using satellite rainfall data combined with qualitative data sources and agent-based modeling to understand how rainfall variability affects migration as well as food and livelihood security in South and Southeast Asia, Sub-Saharan Africa and Latin America [41]. In Stockholm, GPS-equipped vehicles provided real-time traffic assessments and, when combined with other data sets such as those pertaining to weather, made traffic predictions. Such analyses inform urban planning and

also can increase time and cost savings for drivers [42].

**Internet | Search Queries.** The internet stores a vast amount of information, much of which is unstructured. Search queries present one source of data on the internet. In this vein, Google searches for “unemployment” were, found, for example, to correlate with actual unemployment data. Similar data was used to notice changes in the Swine Flu epidemic roughly two weeks before official US Centers for Disease Control and Prevention data sources reflected it [42]. The Bank of England uses search queries related to property, for example, to infer housing price changes [3]. Colombia’s Ministry of Finance uses the information generated by Google searches to assess short-term GDP trends in Colombia and publish monthly macroeconomic reports which discuss the results of the model developed [43].

**Internet | Text.** Text analysis is critical for data generated via the internet not only for sentiment analysis (e.g. favorable/unfavorable views on a policy) but also for lexical analysis to understand elements of culture. One group analyzed the concept of honor in the Middle East, for example, and found how it differed by region and changed over time in response to the events of September 11th. Such analysis could inform the appropriate selection of language in, say, diplomacy or educational materials. Further applications in this regard could include, for example, developing a contextual lexicon on financial literacy in order to tailor microlending by region [44]. By combining topic modeling methods--whereby one explores and understands concepts and topics from text--with sentiment analysis, one can gain a richer understanding of unstructured text data [24].

**Social Media | Tweets.** Similar to the example of analyzing search queries above, social media data such as Twitter tweets can be used as an early indicator of an unemployment hike or to evaluate crisis-related stress [32]. Another case utilized tweets to know about a cholera outbreak in Haiti up to two weeks prior to official statistics [42]. Both of these cases demonstrate the ability to reduce reaction time and improve process with which to deal with various crises. Tweets have been used by hedge fund

managers to make better stock exchange decisions and by researchers to predict a film’s success at the box office or a person’s likelihood to get flu shots [3].

**Financial | Credit Card Transactions.** Credit card companies have increasingly been using their massive stores of data to enhance their services. In several cases, companies use purchase data to identify unusual behavior in real time and quickly address potential credit card fraud [3]. In other cases, financial institutions have been cited as being able to predict whether someone is dating [45] or even infer the strength of a marriage [46].

## Big Data for Development in Central America: World Bank pilot efforts

Since 2014 the World Bank has been exploring the potential utility of big data to support public policies in Central American countries. The starting point was addressing data availability issues, in a context where traditional data collection methods, such as household surveys, are undertaken with a relatively low frequency in Central America and incur high costs. Therefore, the goal was to explore the potential of alternative data sources, such as those one described in the paragraphs above, to fill a data gap. With this objective, three different exploratory pilots focusing on different sources of information (internet data, social network data, and satellite data) were developed.

The objective of the first pilot was to assess the possibility of using web search keyword data (from Google Trends) for nowcasting price series in Central America. The study, which focused on Costa Rica, El Salvador, and Honduras, highlighted the challenges in using Google Trends data. The findings, based on a number of indexes constructed to summarize Google Trends data, showed that Google Trends data can improve the ability to forecast certain price series (especially in Costa Rica and El Salvador, where the web search data was of higher quality).

The second pilot, jointly carried out with the United Nations initiative working on big data (UN Global Pulse), explored the potential of social network content to analyze public perception of a policy

reform in Central America. The project focused on the gas subsidy reform in El Salvador and consisted of gathering data from Twitter. After geo-referencing on-line content to the country and categorizing information based on the content, the study used text analytics to see if the results from the social media analysis closely followed the public opinion as measured through a series of household surveys conducted in the El Salvador before and after the reform. By undertaking what can be thought of as a replication study the goal was to establish the validity of the alternative method (social media text analysis) to capture the underlying phenomenon under study. Preliminary results confirmed that Twitter data provides a useful complement to analyze the public perception of a policy reform.

The third pilot tried to use satellite data to understand poverty levels in Nicaragua and Guatemala. In particular, the objective of the analysis was to produce a first assessment of the information content of night-time illumination measures and explore correlations with poverty at high levels of geographical disaggregation. The analysis showed that the one-to-one correlation is negative and statistically significant, indicating that night-time illumination data may contain information relevant for analyzing poverty conditions.

These pilots are just a starting point. The World Bank launched a Big Data Innovation Challenge in September 2014 to promote big data driven internal projects. In less than a month, more than 130 project proposals were submitted to the Challenge to keep exploring the potential of big data for development.

## Examples by medium and purpose

	AWARENESS	UNDERSTANDING	FORECASTING
MOBILE	<p>A study in Afghanistan has shown that you can use CDR data to detect impacts from “microviolence” like skirmishes and IEDs. Microviolence has clear effects on the ways people communicate and patterns of mobility and migration, similar to what you might see after a natural disaster. [22]</p>	<p>A study in the UK used mobile and census socioeconomic data to examine the connection between the diversity of social networks and socioeconomic opportunity and wellbeing, validating an assumption in network science previously untested at the population level—that greater diversity of ties provides greater access to social and economic opportunities. [47]</p>	<p>Research has shown that when mobile operators see airtime top-off amounts shrinking in a certain area, it tends to indicate a loss of income in the resident population. Such information might indicate increased economic distress before that data shows up in official indicators. [36]</p>
FINANCIAL	<p>Xoom, a company specializing in international money transfers, noticed in 2011 that there were more frequent than usual payments being funded by Discover credit cards originating in New Jersey. All looked legitimate, but it was a pattern where one should not have existed. Further investigation revealed the fraudulent activity of a criminal group. [3]</p>	<p>The Oversea-Chinese Banking Corporation (OCBC) increased understanding of individual customer preferences by analyzing historic customer data, then designed an event-based marketing strategy focused on using a large volume of coordinated and personalized marketing messages. Their precise targeting positively impacted numerous key performance metrics and increased campaign revenues by over 400%. [48]</p>	<p>Predictive analytics tools like FlexEdge allow traders on US equity markets to engage in advanced forecasting, including overnight and intraday forecasts updated by the minute, resulting in an error reduction of up to 25% over standard forecasting techniques which typically take a historical window average [49].</p>
SATELLITE	<p>Following the 2013 typhoon in the Philippines, Tomnod (now DigitalGlobe) took their high-resolution satellite images, divided them into pieces and then shared them publicly to crowdsource identification of features of interest and enable rapid assessment of the situation on the ground: where buildings were damaged, where debris was located, and where roads were impassable. First responders used maps generated through this system and the Red Cross relied on the data to determine resources. The Philippine government also will analyze the data to better prepare for the future. [50]</p>	<p>The Open Data for Resilience Initiative fosters the provision and analysis of data from climate scientists, local governments and communities to reduce the impact of natural disasters by empowering decision-makers in 25 primarily developing countries with better information on where and how to build safer schools, how to insure farmers against drought, and how to protect coastal cities against future climate impacts, among other intelligence. [2]</p>	<p>AWhere’s “Mosquito Abatement Decision Information System (MADIS)” crunches petabytes of satellite data imagery to locate the spectral signature of water primed for breeding mosquitoes and combines it with location intelligence algorithms and models of weather and mosquito biology to identify nascent outbreaks of mosquitoes even before they hatch. [46]</p>

	AWARENESS	UNDERSTANDING	FORECASTING
<b>INTERNET</b>	Pricestats uses software to crawl the internet daily and collect prices on products from thousands of online retailers, enabling them to calculate daily inflation statistics which are used by academic partners to conduct economic research and public institutions to improve public policy decision-making and anticipate commodity shocks on vulnerable populations. [3], [51]	Logawi engaged in a research project using lexical analysis--the use of the internet to create a cultural context for particular words and phrases enabling deeper understanding of how cultures view particular ideas--to assess how different populations across the Middle East understood the concept of "honor." Based on interviews and analysis of internet data, Logawi developed a lexicon of words and phrases that mapped onto the region viewing how definitions and use of "honor" change for different cultures over time. [44]	Research has shown that trends in increasing or decreasing volumes of housing-related search queries in Google are a more accurate predictor of house sales in the next quarter than the forecasts of real estate economists. [9]
<b>SOCIAL MEDIA</b>	Using social media analytics in Syria, SecDev Group was able to identify the locations of ceasefire violations or regime deployments within 15 minutes after they took place, enabling them to rapidly inform UN monitors ensuring swift response. [52]	A project by UNICEF used social media monitoring tools to track parents' attitudes towards vaccination in Eastern Europe by identifying patterns in the sentiments of their public posts on blogs and social media. The study increased understanding of how to respond to vaccine hesitancy and educate parents' to make informed choices, including engagement strategies and messaging. [53]	A collaborative research project between Global Pulse and the SAS Institute analyzing unemployment through the lens of social media in the US and Ireland revealed that the increases in the volume of employment-related conversations on public blogs, online forums and news in Ireland which were characterized by the sentiment "confusion" show up three months before official increases in unemployment, while in the US conversations about the loss of housing increased two months after unemployment spikes. [10]

## Areas of high potential for big data

A variety of authors and institutions have pointed out what they see as areas of high potential for big data. At a broad level, many authors emphasize the potential of combining datasets to enhance understanding [54], [55]. The OECD points to four broad international research topic areas which would benefit from a variety of data types. These topic areas include population dynamics and societal change; public health risks; economic growth, innovation, research and development activity; as well as social and environmental vulnerability and resilience [55].

Beyond research, however, there is a need for more specific, practical arenas within which big data shows promise. The United Nations' Global Pulse argues

that global development work can be improved using big data in three ways: strengthening early warning systems to shorten crisis response times, enhancing awareness of situations on the ground to better design programs and policies, and enabling real-time feedback to make appropriate and timely adjustments [10]. These categories are examined below. In addition to these areas, however, several individuals have highlighted the promise that big data shows in terms of strengthening understanding around complex systems dynamics thereby enabling better policy-making. To the extent that specific challenges are elucidated and data is used more in the context of ongoing processes rather than one-time projects, big data will have stronger impacts on international development.

## ***Early warning***

Various other sources have also emphasized the potential of big data for early warning [54], [30]. Two concrete examples of such work include forecasting riots using food price data with relevant proxies [56], or predictive policing whereby various data sources, including police databases, are combined with mathematical tools to route police patrols in anticipation of future crime [57].

## ***Enhancing awareness and enabling real-time feedback***

The potential of big data for enhancing real-time awareness, also known as nowcasting, is also repeatedly discussed by other individuals [37], [12]. In fact, approximately 1,200 business and IT professionals attributed “real-time information” as one of the top three defining characteristics of big data [58]. Several examples testify to the power of using big data to enhance awareness. Two MIT economists, for example, used internet data to collect prices on half a million products and detected price deflation two months prior to release of the official, and expensive-to-produce Consumer Price Index [3]. Alternatively, after a recent tsunami in Japan, Honda was able to provide road closure data within a day using GPS data from newer generation cars [45]. One idea discussed among several big data experts and practitioners included creating publicly accessible databases to enable anyone to assess financial markets, thereby protecting consumers and investors [37].

## ***Understanding and interacting with social systems***

One interviewee discussed the possibility of studying the growth of urban boundaries, such as favela growth in Brazil using historical satellite data combined with complex systems modeling. This could lead to understanding city growth patterns and improved city and regional planning [59]. A few authors have also discussed the possibilities of opening big data as well as relevant analytical capabilities to level the playing field for labor and/or product supply from a variety of sources [39], [60]. With larger-scale players dominating

the retail market, sourcing is often simplified by working with large scale suppliers, crowding out smaller producers. A third-party organization could, however, utilize big data analytics to ensure replenishment and coordinate supply from a variety of product sources, large or small [39].

## ***Specific challenges and ongoing processes***

To move the agenda of big data for development forward, more than general categories or approaches will be needed. Reiterated during one interview [48] is that what is especially needed in the development big data space is the specification of challenges that lend themselves to the utilization of big data. Put another way, the kinds of insights that need to be generated should be specified, and such a specification process would benefit from the input of practitioners, data scientists, designers, and relevant thought leaders. Once specified, a space can be created for those with the necessary contextual and analytical ability to propose methods to address the carefully defined challenges using big data. The World Bank, in collaboration with other institutions and organizations, may play a crucial role in this regard as a convener of various parties both to specify challenges and explore ways to address them [23], [24]. By convening parties, a shared way of speaking and thinking about big data can be created which is general enough to be inclusive of a diversity of approaches yet specific enough to generate meaningful action.

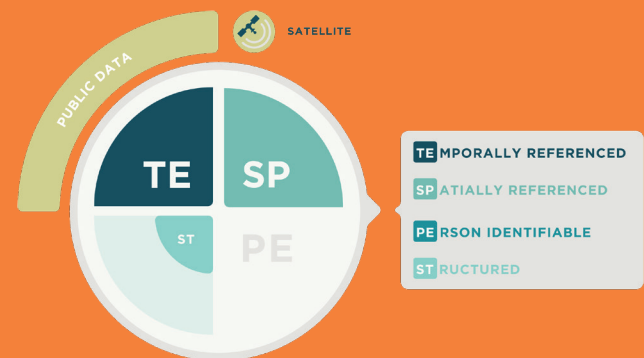
Furthermore, while short-term projects using big data can be helpful to increase awareness or begin to adjust systems to be more effective, the value of big data is perhaps most evident when it is integrated into ongoing processes. Examples of such process-oriented uses of big data range from private sector retailers using big data to minimize inventories to public sector governments streamlining tax collection mechanisms, unemployment services, or city emergency services. Indeed, if big data will be used to address challenges, it will have to be integrated into ongoing processes. Only in this way can its use be refined over time and the necessary knowledge be generated more effectively to improve those systems of interest.

# CASE STUDY

# DATA PALETTE

## Forecasting and Awareness of Weather Patterns Using Satellite Data

Weather and land conditions via remote sensing imagery



### Motivation

Heavy rainfall in the city of Rio de Janeiro often leads to severe landslides and flooding causing significant public safety issues. For rescue efforts, coordination is needed among several different emergency agencies when this happens. On April 6, 2010 the city had one of its worst storms, with mudslides leaving over 200 people dead and thousands without a home. It was this event along with the fact that the city was preparing for the 2014 World Cup and the 2016 Summer Olympics that pushed the city to use data and predictive modeling for emergency response.

### Data Generation

As indicated by the data palette above, the data used to predict weather forecasts is primarily satellite data gathered from open sources such as the National Oceanic Atmospheric Administration (NOAA) while sea surface temperatures are collected directly from NASA. For predicting landslides and flooding, data is pulled from the river basin, topographic surveys, the municipality's historical rainfall logs, and radar feeds. Data such as temperatures at different altitudes, wind data, and soil absorption is captured to help develop accurate predictions. The city is also using loop detectors in the street, GPS data, and video data to help plan emergency responses. As indicated in the palette above, the weather pattern data is not highly structured, however it is highly spatially and temporally referenced as it provides frequent and specific geographic information representing a state in time.

### Data Interpretation

Data from across 30 different city agencies is all housed in the Rio Operations Center providing a holistic perspective of how the city is operating. The city created the Rio Operations Center to allow for real-time decision making for emergency responsiveness and to improve safety based on consolidated data from various urban systems. The idea behind creating this operations center was to remove silos between the different emergency response services (i.e. police department, firefighters, etc.).

Data is analyzed using a variety of algorithms that allows for projections of floods and landslides to be made on a half-kilometer basis and is able to predict heavy rains up to 48 hours in advance. The basic model that is used for predicting heavy rainfalls is IBM's Watson Weather Model. This model has been configured to the city of Rio based on a comprehensive list of weather related events that was given to IBM by the city.



A combination of real-time and historical data is currently being used for this analysis. Rio has very good historical flooding data available cataloguing at least 232 recurrent points of flooding. The algorithms are very granular, taking raw data about wind currents, altitude temperatures, humidity level of the soil, and the geography of the city to create accurate predictions of landslides. This data is then analyzed against the infrastructure of the city to determine the likelihood of floods and landslides. For example, rainfall predictions are compared against the layout of city infrastructure, the number of trees that can help absorb some of the water and the conditions of the soil to predict greater risk areas for floods.

The city uses IBM's Intelligent Operations center that pulls in information from multiple sources and provides an executive dashboard for city officials to quickly gain insight into what is happening across Rio in real time. City officials are able to see high-resolution animations of two and three-dimensional visualizations of key weather variables and obtain detailed tables of weather data at various locations.

## Insight Implementation

The new alert system notifies city officials and emergency responders in real time via automated email notifications and instant messaging. As a result of their high-resolution weather forecasting and hydrological modeling systems, Rio has improved emergency response time by 30%. An additional benefit of the new alert system is all of the data that it generates from the receipt of a message to the response taken. Analysis of this data allows city responders to improve their current procedures resulting in lower response times and greater coordination of activities.

## From Idea to Ongoing Process

The Rio Operations Center was the first center in the world to integrate all stages of disaster management from prediction, mitigation and preparedness, to immediate response and feedback capture for future incidents. By having access to each other's data in a non-siloed environment, greater communication and collaboration was seen between emergency response services leading to more rapid response times. In addition to being able to predict rain and flash floods, the city is now able to also assess the effects of weather on city traffic and power outages by using a unified mathematical model of Rio. Moreover, Rio is now going beyond weather forecasting to integrate big data into other areas of municipal management. For example, data on waste collection via GPS systems installed in trucks is also collected by the Rio Operations Center to create a comprehensive picture of how public services are operating in the city.

The city of Rio has also made this data publicly available for its citizens to be able to better manage their lives. The mayor of Rio de Janeiro, Eduardo Paes stated "in addition to using all information available for municipal management, we share that data with the population on mobile devices and social networks, so as to empower them with initiatives that can contribute to an improved flow of city operations". Citizens can receive daily data feeds by following the Rio Operations Center updates on Twitter @OperacoesRio and Facebook at Centro de Operações Rio. These sites also provide recommendations for alternative routes during special events as well as current traffic and weather conditions. Eduardo Paes, the mayor of Rio has stated that by having these updates, the quality of life in Rio has increased significantly and that this is helping bring more businesses and people to Rio.

Rio is using big data as part of its daily operations to transition to being a smarter city and providing a higher quality of life for its citizens. In fact, the mayor of Rio has made technology part of its "4 commandments for smarter cities" stating that "a city of the future has to use technology to be present". Rio has continued its

partnership with IBM to continuously improve upon its original algorithms so that as technology advances, the city is also able to stay ahead of the game. As a result of this pilot, Rio is now fully committed to using technology as a way to help govern the city.

## References

Hilbert, M. (2013). Big Data for Development: From Information-to Knowledge Societies. Available at SSRN 2205145.

Conversation with Jean-Francois Barsoum from IBM, March 2014.

Conversation with Renato de Gusmao from IBM, March 2014.

Treinish, Loyd. (2014). Operational Forecasting of Severe Flooding Events in Rio de Janeiro. Retrieved from: <http://hepex.irstea.fr/operational-forecasting-of-severe-flooding-events-in-rio-de-janeiro/>

IBM. (2011). City of Rio de Janeiro and IBM Collaborate to Advance Emergency Response System; Access to Real-Time Information Empowers Citizens. Retrieved from: <http://www-03.ibm.com/press/us/en/pressrelease/35945.wss>

Eduardo Paes TED Talk. (2012). The 4 Commandments of Cities. Retrieved from: [http://www.ted.com/talks/eduardo\\_paes\\_the\\_4\\_commandments\\_of\\_cities](http://www.ted.com/talks/eduardo_paes_the_4_commandments_of_cities)



NOKIA

+919848923623

:  
1) Rainfall: Light  
to moderate

Options Back



ए.औं  
def 3

क.ख  
jkl

ट.ण

## SECTION 4

# HOW CAN WE WORK WITH BIG DATA?

Technology alone is not sufficient to understand and interpret results from use of big data. Turning big data into insights which are then acted upon requires an effective combination of both technological and human capabilities.

### Technological capabilities

Each phase of the data interpretation process discussed above highlights the technological capabilities necessary to work with big data effectively. First, in terms of accessing data, it is important to have the necessary hardware and software to collect data depending on whether a dynamic or static method is utilized. If data is dynamically fed from an online source such as Twitter, for example, then the analysis software must allow for such real-time, continuous updated analysis. If, instead, data is being downloaded from some source and then kept for later analysis, it is important to ensure sufficient hardware capacity to store such data.

Given sufficient technological capacity simply to access data from various sources, it is necessary to have software and hardware capabilities to connect and interact with large and diverse datasets. The larger the datasets, the greater the hardware storage and processing power and the more scalable a software platform are needed through which to process queries and pull data for analysis. With large-scale analyses, software is often needed to make use of parallel computing through which computations of massive amounts of data can be processed simultaneously over multiple processors [6]. Relational

database software such as Microsoft Access, for example, will scale very poorly in the face of dozens of gigabytes, let alone terabytes of data. Instead of using supercomputers, software may be used to conduct parallel data processing over multiple computers, including even videogame consoles, more cheaply. Examples of such software include Hadoop clusters and products such as Microsoft's Azure, and Amazon's EC2 [42], [6]. Open source tools, such as those used by the city of Chicago's predictive analytics platform, which utilizes big data [31], present more financially inexpensive software options for analysis. Altogether, these cheaper parallel processing and analysis options are promising when considering the lack of big data hardware and software capacity in many developing countries [42].

The more diverse the datasets of interest, the more robust a software platform must be through which to interact datasets. For example, in the case of structured column/row datasets, interacting datasets may be as simple as identifying unique keys to join tables, as is done using relational database software like Microsoft Access. However, when considering relatively unstructured data such as satellite infrared data or a collection of hundreds of millions of strings of text, software capabilities are critical in order to analyze effectively and connect them to structured datasets for generation of appropriately formulated insights.

Beyond software and hardware requirements, it is immensely helpful when the data which is utilized has appropriately encoded metadata, i.e. data which

describes each dataset. In particular, Global Pulse recommends that metadata describe the “type of information contained in the data”, “the observer or reporter”, “the channel through which the data was acquired”, “whether the data is quantitative or qualitative,” and “the spatio-temporal granularity of the data, i.e. the level of geographic disaggregation (province, village, or household) and the interval at which data is collected” [32]. Given such metadata, analysts and decision-makers can more easily identify the provenance of a particular dataset. This is especially helpful when analyzing data mashups, or interactions of multiple datasets. With complete metadata, in other words, the analysis is more transparent regarding assumptions.

Beyond simply being aware of the data sources via metadata, some authors have highlighted the need for analysts and decision-makers to understand more effectively the assumptions of the model and/or combined dataset by, in essence “playing” with the assumptions. In this regard, several authors have explored the concept of a hypothetical Analytics Cloud Environment through which a user can change assumptions and see their impact on an analysis [6]. In designing such software to be scalable, well-designed Application Programming Interfaces (APIs) must be created to channel data at an optimal level of aggregation, so that users may fluidly interact with a large database [25].

## Human capabilities and data intermediaries

A reading of the above technological capabilities also indicates the undoubted necessity for human capabilities to interact meaningfully with big data. The need for human capacity to understand and use big data effectively, particularly in government and the public sector [12] [5], or specifically in developing countries [42], is reiterated by various authors and agencies [10], [33], [36], [61], [40]. At a fundamental level, to work with big data requires a shift in mindset from working with “small” data. Some authors emphasize that this implies the ability to analyze vast amounts of information rather than

only samples, a willingness to work with messiness, and an appreciation for correlation rather than a strict interest in causation [3]. A senior statistician at Google pointed out that a good data scientist needs to have computer science and math skills as well as a “deep, wide-ranging curiosity, is innovative and is guided by experience as well as data” [62]. Other necessary skills include the ability to clean and organize large data sets, particularly those that are unstructured, and to be able to communicate insights in actionable language [11].

In its report on big data, McKinsey points out, however, that a “significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data” [16]. Indeed, the human capabilities needed are wide-ranging from those having to do with technology to those related to people and real-world problems, e.g. hardware setup, data management, software development, mathematical and statistical analysis, real-world model development and assessment, as well as the distillation and communication of actionable insights. Beyond such skills, an intimate knowledge of the real world situation of interest is critical [40].

Beyond the individual skills and capacities required, effective spaces and environments need to be created in order for multiple viewpoints to advance the analysis collaboratively. At one level, distributed labor coordination mechanisms such as crowdsourcing can be utilized to aggregate thousands or even millions of people’s perspectives on a dataset in order to complement and strengthen the big data analysis [50]. At a smaller, yet more complex level, collaborative environments need to be created through which the diverse perspectives of those working with big data can come together to produce new, more comprehensive insights.

Given the required individual and collective capacities to work with big data, it is no surprise that in surveying thousands of businesses around the world,

six out of ten said that their “organization has more data than it can use effectively” and the leading obstacle was a “lack of understanding of how to use analytics to improve the business.” The majority of these businesses which frequently used data analytics actually used a “centralized enterprise unit” which

served as a space in which the right combination of skill sets among individuals could come together to do big data analysis [17]. These enterprise units are examples of data intermediaries that will undoubtedly be needed in the coming years to make sense of big data [37].

## Volunteer Technical Communities

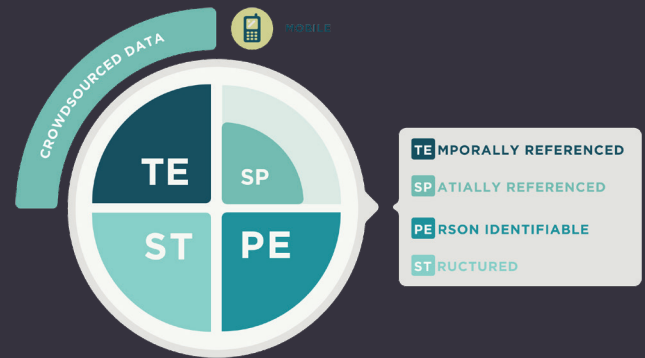
Governmental or non-governmental institutions that wish to transform raw datasets into practical tools often lack the expertise to do so. Yet this should not be a limiting factor. One avenue which is proving itself valuable to leverage the skill sets of individuals outside of an organization is volunteer technical communities such as hackathons. In these settings, subject matter experts, technologists, designers, and a variety of other participants, come together to define problems, share relevant datasets, and rapidly prototype solutions. Often these events are followed up by acceleration phases in which the most promising solutions are further developed. Examples of hackathons that have been used in an international development context include the Domestic Violence Hackathon held in Washington D.C. as well as countries in Central America. Another example is Random Hacks of Kindness which, between 2009 and 2013 organized hundreds of community-focused hackathons around the world and engaged in a similar process. These volunteers created, for example, InaSAFE, a web-based tool which combines several relevant datasets and strengthens decision making around natural disaster response [63].

## Connected Farmer Alliance

Crowdsourced supplier data via mobile phones

### Motivation

Vodafone has a deep interest in development in emerging markets, particularly in Africa where their mobile financial solutions like M-PESA (a mobile money transfer platform) and M-Shwari (a partnership between M-PESA and CBA to provide a savings and loans service) have a strong presence. Vodafone's interest in development and the pursuit of disruptive innovation tied with a clear potential for commercial businesses to play a role in supporting the agricultural sector (an area of focus for many of Vodafone's African enterprise clients) led it to join with USAID and TechnoServe to form the Connected Farmer Alliance. This Alliance pilots initiatives aiming to create a better ecosystem for mobile services in the agricultural sector, impacting production through the supply chain to enterprise use.



### Data Generation

The program focuses on Kenya, Tanzania and Mozambique, and is divided into 3 distinct areas of focus: enterprise solutions to source from small farmers, improving mobile financial services and mobile value-added services. The first area, where much of the testing has already taken place, involves enterprise solutions which enable enterprises to better source from small farmers and allow farmers better access to markets. The data is gathered and distributed through a suite of modules, including a registration module allowing an agent of an enterprise to register farmer (or for farmers to register themselves as suppliers) who supply a particular produce. The service enables a remote crowdsourced data-gathering method to identify who and where farmers are and the crops they specialize in producing. The data gathered through mobile phones in this module is highly structured and referenced both temporally and spatially, as well as highly person identifiable, enabling enterprise participants to distinguish specific farmers and their products. The typical enterprise participants are mid-sized national companies who source their produce from small farmers and are seeking more detailed data and interaction with available suppliers. Building upon the crowdsourced supplier data are a series of additional modules including two-way communication that enables enterprises to share information with, or survey, farmers. A receipting module, integrated with M-PESA, allows enterprises to send receipts and pay farmers at the point of sale, identifying volume of purchase, time and price and increasing transparency. Another module allows enterprises to offer short-term loans through M-PESA, enabling cash advances that are later deducted from payment for produce. Finally a tracking module enables enterprises to better track collection processes and points to streamline product collection. At the pilot phase the size of the crowdsourced dataset does not yet approach big data, however Vodafone is currently preparing to bring this first suite of modules to commercial markets for much broader deployment.

The second area of focus, currently in the conceptual development and partnership-building phase, involves the improvement of mobile financial services. One area of research is the extent to which big datasets of historical mobile financial transactions, generated through other Vodafone products and services, can prove useful in assessing credit-worthiness of loan applicants. This product area may also work with local partners, to incorporate

the use of mobile financial data in streamlining insurance pricing and payouts for farmers by using location data to assist insurers in more rapid analysis of claims.

The third focus area, only at the earliest conceptual stages, is to use the enterprise solutions and mobile financial services created in the first two stages of the product to create a supportive environment overall for mobile value-added services for anyone wanting to take products to market. This area will also include the growth of business development and incubation services to sustainable mobile business growth in the agriculture sector.

## Data Interpretation

Vodafone works with subsidiary Mezzanine on the development and management of the data collection platform, which is locally hosted in the Kenyan, Tanzanian and Mozambican markets themselves and protected by high-level security mechanisms. In pilot phase, data is available only to the enterprise and participating farmers and for the surveys, enterprises receive only aggregated responses, not individual records. Vodafone is working with enterprise customers on the most convenient way for farmers to submit data whilst ensuring confidentiality for them and businesses. The details of data privacy will be governed by Vodafone's data privacy policies to ensure ongoing protection.

Within the Connected Farmer Alliance partnership, TechnoServe is charged with analysis and interpretation of how the modules are performing for the enterprises and farmers. However given the small sample set involved in the pilot of the enterprise modules, insights are currently being gathered through traditional survey methods. Those methods include assessing goals for the participants at the project outset, determining areas of measurement, and collecting input through questionnaires during the process. Additionally, the Connected Farmer Alliance supports enterprise partners in their own data analyses of information and outcomes.

## Insights for Action

Although the project is still in its early phases, insights are beginning to emerge around the benefits of the enterprise modules. Cost savings have been shown on the part of farmers who receive M-PESA for loans and payments. By receiving M-PESA, these farmers avoid costly, time-consuming, and risky trips to the enterprise office to collect cash. The receipting module has resulted in cost savings for enterprises due to operational efficiencies and improved process transparency. A key benefit of mobile solutions for farmers is an increase in access to information. Nonetheless it is difficult to make generic content services meaningful to small farmers whose local realities may vary significantly within a distance of just a few kilometers. The targeted information flow permitted by the two-way information module has been shown to provide information particularly relevant to the stakeholder farmers, as well as to enhance face to face interactions among farmers and enterprises.

## From Idea to Ongoing Process

Although the Connected Farmer Alliance has a clear social transformation element which inspires the partners and enterprises alike, growth and use of these mobile tools and long-term sustainability of the piloted approach will fall under Vodafone's commercial initiatives. With the specific intent of going beyond the pilot phase and putting in place publicly accessible tools, Vodafone is currently in the process of scaling up the mobile tools used in the first phase of the Connected Farmer Alliance project for commercial use as a method to generate large scale, targeted and valuable data for small farmers and enterprises alike.



## References

Conversation with Laura Crow, Principal Product Manager for M-PESA, March 2014.

Correspondence with Drew Johnson, Program Manager TechnoServe Tanzania, June 2014

TechnoServe. (2014) Projects: Connected Farmer Alliance. Retrieved from: <http://www.technoserve.org/our-work/projects/connected-farmer-alliance>



## SECTION 5

# WHAT ARE SOME OF THE CHALLENGES AND CONSIDERATIONS WHEN WORKING WITH BIG DATA?

As the data deluge continues to grow and forward thinking managers and policy makers seek to make use of it, challenges at the levels of expertise, appropriate use, and institutional arrangements come to the forefront. Whereas in the past, smaller-scale, and less diverse datasets could be analyzed by a limited number of individuals, big data requires a wider skill set. An added dimension to the use of bigger data is its ability to understand and predict ever-smaller segments of the population, sometimes even to the level of the individual. The ability to be aware and even forecast the behaviors of such small segments, including individuals, raises new ethical questions. Finally, whereas data analysis was often restricted to that collected by a single institution, the nature of big data requires new forms of inter-institutional relationships in order to leverage data resources, human talent, and decision-making capacity.

The following sections organize challenges and considerations according to stages of the Data for Action Framework [Figure 2] discussed in section two. First a series of considerations is discussed with respect to the various ways in which data is generated, stored, and accessed. Then challenges around how effectively to manage and analyze the data are enumerated. Along similar lines, practical considerations arise when discussing how insights are actually defined. Cultural and ethical challenges then come to the forefront when considering how to actually implement insights.

### Data generation process and structure

Several challenges must be overcome and considerations kept in mind regarding the data generating process and the data structure itself.

To begin, the very trigger which encodes behaviors into data can have implications for analysis. If data is passively recorded, then it is less susceptible to the statistical problem of selection bias, i.e. the data that is collected is systematically unrepresentative of the behavior of interest. If instead the data is actively selected, then it is more susceptible to such a bias. If an individual is interested in collecting her walking data throughout a week, she may, for example, input data into a spreadsheet on those days when she remembers. This may, however paint a biased picture of her movement since she only records data when she walks great distances, therefore biasing the report of her movement. If, instead of collecting data actively, she used a wristband which passively collected data, then a more representative picture would be drawn.

Once encoded, it is important to consider the features of the datasets of interest. If they are unstructured, for example, they will require the development of appropriate processing methods, especially when integrated with structured data. Mobile data, like that analyzed in the case study on Understanding Labor Market Shocks using Mobile Phone Data, is received in unstructured repositories of millions of individual files, requiring time-intensive processing, programming and the use of expensive hardware to obtain indicators of population

movement and prepare for interaction with other data [22]. Text analysis is one example of making sense of, say, unstructured Facebook status updates or online Tweets. The method used to structure unstructured data adds yet another point at which decisions are made, making the analysis further susceptible to biases that the researcher may have.

Whether data sets have methods to identify temporal, geographical, or individual characteristics, such as time/date stamps, latitude/longitude information, or unique personal IDs, respectively, will determine to what extent data mashups are possible. However, a challenge that must be addressed in combining such data sets is to ensure proper aggregation. For example, if one data set is collected every minute, but another is collected every day, then, to ensure comparability, the analyst must carefully consider how to aggregate the minute-based data so that it can be effectively joined with the daily data.

The media through which behaviors are encoded into data each present their own series of challenges. Some of these challenges are inherent to the medium, while others are due to association with certain features of the data structure or generating process. Mobile phone data, by their very nature, are highly sensitive and should be treated carefully. Although mobile data is highly disaggregated and can be very rich, it has been observed that its analysis should be validated through corroboration with other sources such as household surveys or satellite data [40]. Satellite data can become very large very fast, since it is primarily high-resolution image data. It can also be highly unstructured, particularly when it comes to visual pattern analysis, and may benefit especially from human review. Social media data is often in the form of unstructured text which requires specific analytical capabilities to codify and identify useful patterns. One researcher identified techniques such as topic modeling and named entity recognition to be useful in this regard. In the case of the Global Pulse program tracking food price inflation in Indonesia through Twitter, a researcher trained a sentiment classification algorithm by manually classifying Tweets according to various sentiments, allowing the algorithm to, in turn, classify other Tweets automatically [64]. The benefit of beginning

with exploratory, visual analysis of social media in order to see patterns and build a formal statistical model was also noted [24]. Social media often store actively generated data such as Twitter Tweets and may therefore suffer from selection bias. On a related note, retweets and sharing of links is actually the bulk of Twitter traffic, such that a very small minority control the agenda of what is originally created. This presents challenges of ensuring that analyses which use social media place results in the right population context [45]. Other information found on the internet such as webpages, blogs, videos, etc. may share similar problems as those noted above regarding identifying the accurate population being represented as well as effectively interacting with unstructured data. Point-of-sale or internet sales data are often high-frequency and high-volume datasets which require effective structures to process massive stores of data constantly as well as thoughtfully constructed models which appropriately consider the relevant, short-run nature of the decisions being made by humans [39].

## Data interpretation process

To access, prepare, and analyze data sets effectively presents a series of institutional and technical challenges. Beginning with access, several institutional challenges must be overcome just to enable the ready sharing of data sets. Once data is accessible, several technical and data management challenges must be overcome.

### Access

One of the first challenges which must be overcome in order to conduct big data analyses is for data to be more openly shared [65]. This is particularly important for those development-oriented institutions which do not, themselves, generate the data of interest. Indeed, one data scientist/artist pointed out how one of the biggest challenges he faces is simply trying to access data from institutions, including government agencies, which hold vested interests and/or a commodity-ownership perspective over the data they store [25].

An entire ecosystem is needed to open and use big data effectively [16], [36]. Common data standards and

sharing incentives constitute two aspects of such an ecosystem. Leading international agencies will have to address the challenge of collaborating to define and agree on efficient and well-coordinated sharing mechanisms [55]. Standards for data integration, such as APIs, are needed, as are standards to coordinate data generation. Examples of mechanisms to develop both kinds of standards include IEEE, IEC for smart grid, or the Global Earthquake Model for disaster preparedness.

As the challenge of shared standards is overcome, the incentive to share the data must be strengthened. To this end, business models need to be developed to ensure that private sectors are willing to share data [36]. Also, governments need to design policy to help capture the value of big data and enable sharing across agencies [5]. A particular challenge in this regard is the definition of intellectual property rights which retain data ownership yet allow researchers and/or decision makers to use the data for perhaps originally unintended purposes [26]. In addition, governments may have to consider the design of privacy policies for personal and proprietary data, safeguards for information misuse, and regulations on financial transactions and online purchases [12].

As the challenge of opening data in the last decade has partly been addressed, several nascent phenomena have emerged. Beyond enabling analysis, opening data and making it available to the public motivates citizens to engage personally with the data and, in some cases, correct information and improve the accuracy of government databases [5]. Furthermore, opening data can serve as a catalyst for engaging large groups of citizens to apply their technical capacities and desire for social betterment to design novel ways to leverage data for the public good (e.g. National Day of Civic Hacking, or the International Space Apps Challenge). Opening data has also contributed to the emergence of an open science paradigm among academics concerned with enabling research accessibility by the public, permitting wider use of research findings, facilitating research collaborations, and/or simply sharing datasets with other researchers [66]. An additional challenge, however, that has emerged is the fact that as data is opened and its value is recognized

by private players, they are less willing to share their data. One researcher pointed to the fact that accessing data from companies 10 years ago was much easier, for example, than it is today, partially for this same reason [26]. The researchers in the case study Understanding Labor Market Shocks using Mobile Phone Data indicated the ability to obtain mobile data would be virtually impossible for private sector agencies and extremely challenging even for governments and multilaterals, often relying on personal relationships. Much of the research in that area, therefore, is being done by academia [22].

Whether data is accessed through the public or private institutions presents different challenges. Public institutions can often release data for free; however, administrative hoops, which the above discussion emphasizes, can present great barriers to access. Moreover, as the researchers in the Billion Prices Project discovered, public agencies are often so accustomed to using traditional data sources that they face an additional cultural hurdle to engaging with big data that may take time and use to overcome before its value is internalized [29]. Private institutions, on the other hand, have massive stores of data, the value of which is being increasingly recognized. In this case, security and intellectual property concerns may exist. The Billion Prices Project methodology of scraping price data from retailers addressed the privacy concerns of enterprises-by building in lag time between the data collection and data sharing, and by sharing data at aggregated levels [29].

On the other hand, directly sourcing data from crowds of people, often via an online aggregation mechanism such as one of the many social media tools, presents the unique challenge of ensuring wide and high quality participation. If participation is meager, then the data collected will not only be insufficient from a quantitative standpoint, but the perspective may not be reflective of a broader set of the population. The Global Pulse project profiled as a case study in this report—Tracking Food Price Inflation using Twitter Data—chose to focus on a part of the world where many people Tweet, such that Twitter represents a broader segment of the population, for this very reason [64]. Given that individuals who participate in

# CASE STUDY

## DATA PALETTES

# Tracking Food Price Inflation using Twitter Data

## Motivation

The Global Pulse lab set out to investigate the possibility of utilizing social media data to give an indication of social and/or economic conditions. In particular, they investigated the relationship between food and fuel price Twitter posts and the corresponding changes in official price index measures. By conducting the research in the Indonesian context, the research benefited from a large user base--the city of Jakarta has the largest Twitter presence in the world with 20 million user accounts.

## Data Generation

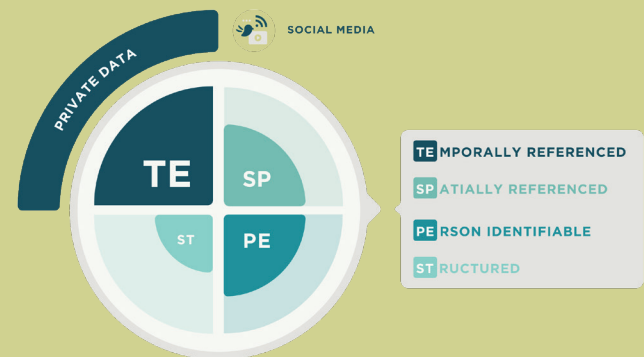
The Twitter data used was generated between March 2011 and April 2013 and formed a largely unstructured dataset of over 100,000 Tweets that were highly temporally referenced, spatially referenced by region and identifiable by Twitter account, and at times by person. This data was complemented by structured public datasets regarding food and fuel prices including official price indices from the Indonesian State Ministry of National Development Planning (BAPPENAS) and the World Food Program (WFP). In particular, CPI data for general foodstuffs came from the Indonesian Office of Statistics (BPS) and data on milk and rice prices from the WFP, both datasets typically generated through questionnaires and surveys. As Indonesia was also experiencing soybean shortages during the period of study leading to the import of soy from the U.S., soybean inflation data from the U.S. was also collected from the World Bank.

## Data Interpretation

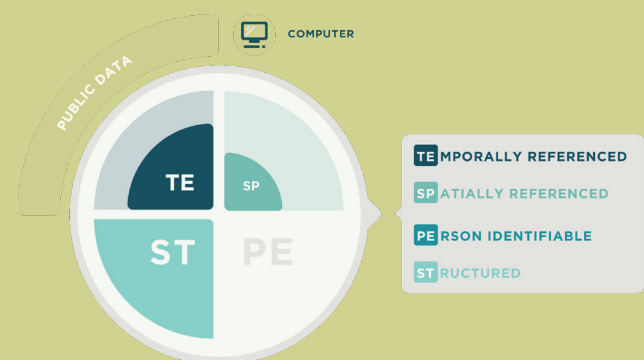
Data from the Twitter "firehose", which streams all tweets live, was captured from March 2011 to April 2013. Full access to the Twitter firehose is difficult to obtain, however Global Pulse was able to secure it through their use of the Crimson Hexagon software, which collected and stored tweets in a database, which could then be analyzed. Other services that could provide similar firehose access include DataSift and Gnip.

The Crimson Hexagon ForSight software includes a classification algorithm, which can analyze strings of text and sort them into categories of interest. For this study, data was categorized through an initial filter of content

Price sentiments posted via Twitter



Official price statistics via public agency surveys



based on keyword searches as being related to food price increases or fuel price increases. Specific words in the Bahasa Indonesia language were utilized as keywords, which the algorithm used to filter those tweets which dealt with the aforementioned price increases. Then, a researcher manually classified randomly selected tweets based on sentiment as “positive”, “negative”, “confused/wondering”, or “realised price high/high-no emotion.” This manual selection by the researcher essentially “trains” the algorithm, which can, in turn, automatically classify the remaining tweets automatically.

By the end of the process, over 100,000 tweets were collected and categorized, forming part of a dataset, which could be analyzed using simple statistical regression techniques. Using such techniques, correlation coefficients could be estimated to analyze the relationship between twitter conversations and official food price inflation data, among other questions.

## Insight Implementation

Some of the final conclusions of the report indicated a relationship between official food inflation statistics and the number of tweets about food price increases. In addition, another relationship was discerned between the topics of food and fuel prices within twitter data. Altogether, this initial effort to analyze social media data indicated the potential to utilize social media data to analyze public sentiment as well as objective economic conditions.

That said, the research demonstrated that, while there was certainly a relationship between the twitter data and official statistics, there was also an abundance of false positive relationships, i.e. large changes in twitter data with no corresponding change in actual inflation measures. More research is certainly needed to improve the classification process as well as the process of geolocation--using arbitrary strings in social media profiles to arrive at exact geographic coordinates--to more fully take advantage of the heterogeneity of social media data and associate sentiment with particular regions of a country. Finally, higher granularity of official statistics are needed in order to more effectively compare it to the correspondingly spatially and temporally specific twitter data.

## From Idea to Ongoing Process

The research has indicated that semi-automatic, retrospective analysis is possible for social media data. To the extent that classification algorithms are strengthened, and more fine grained economic data with which to train algorithms are made available, the potential to implement ongoing real-time analysis of such data appears to be closely within reach.

## References

UN Global Pulse. (2014). Mining Indonesian Tweets to Understand Food Price Crises. <http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf>

Correspondence with Alex Rutherford, Data Scientist, UN Global Pulse, April 2014.

crowd-sourced endeavors vary in their skill, interest, and motivation level, it is possible that mechanisms will need to be put in place to ensure rewarding desired behavior as well as to develop effective quality control processes to check each other's inputs.

## **Preparation**

Once data can be accessed, analysts sometimes consider filtering the quantity and types of data needed. These filters, however, need to be carefully considered; otherwise they may preclude useful information [65]. Regarding data filtration, many big data practitioners estimate that 80% of the effort when working with data is simply cleaning it so that it can be readily analyzed [34]. A critical step in Global Pulse's project tracking food price inflation in Indonesia was the filtration of the data from the Twitter fire hose, for example the researchers had to filter out Tweets in English and Indonesia's numerous local dialects to isolate Tweets in the predominant Bahasa Indonesia language [64].

Once the data is cleaned, scientists must deal with the challenge of how to manage large datasets [65]. In managing these, subjectivity can be introduced when attempting to lend structure to unstructured datasets. With multiple sources, ensuring accurate metadata is critical [65]. When well documented, metadata gives an indication of the provenance of a dataset, thereby increasing transparency and improving understanding of results [37], [54]. With regard to combining multiple data sources, one interviewee warned against using too many all at once. Instead, it is helpful to begin with a couple of relevant data sets and then, as capabilities develop and questions are refined, other datasets can be added [25].

A frequently cited challenge when managing data fed from various companies and/or agencies is ensuring individual privacy and security [36], [65], [12], [32]. For example, although individual-level health or financial data can be used to assist with specifying

appropriate medical treatment or a financial product, consumers may be concerned with sharing such information [5]. Adding to the complexity of this challenge is the fact that each country has different regulations around data privacy [18]. One way to address such a challenge is the development of an internationally recognized code of conduct regarding the use of personal data, which might include best practices regarding data acquisition, sharing, and anonymization<sup>2</sup> [54].

## **Analysis**

Given a well prepared and structured dataset, a series of considerations must be kept in mind. First, large datasets do not always preclude the use of statistical methodology to account for the degree to which a finding is representative of a population. A data scientist will be aware of the various ways in which, for example, posts by Twitter users are not representative of the world's thoughts and opinions [42]. In particular, the selection bias which occurs when inferring real-world behavior using digital sources of information must be kept in mind. In other words, people who use digital sources of information such as blog posts or online shopping, may be systematically unrepresentative of the larger population under consideration. A related concern pertains to the fact that analyzing what has been said about behaviors is different from analyzing behaviors themselves [10].

Also important is that modeling take into account the data generating process. For example, one interviewee pointed out how small price changes derived from retail scanner data may be due to the fact that such prices are actually imputed from data which often comes in the form of weekly total value and quantity sold. Using such imputed prices may fail to take into account systematic relationships underlying the data to, say, mid-week price changes, the use of coupons by certain individuals, etc. [29]

---

2. The question of anonymization itself must be examined carefully. Even when a dataset with personal identifiers--e.g. name or social security number--is randomized, such assignments to individuals may be imputed. One way to deal with this may be simply to aggregate the data sufficiently to enable analysis but to preclude privacy or security concerns [26].



Another consideration when working with big data is the tendency toward apophenia--or seeing patterns when there are in fact none [13]. One interviewee for this report pointed out how, by its very nature, big data can give rise to highly significant statistical correlations. The tendency to assign meaning to these statistically significant relationships must, however, be tempered by rigorous model development [39]. Although the adage which states that correlation is different from causation is critical to understand, and every big data scientist keeps it in mind, the value of correlations remains. For example, correlations may provoke interpretive stories, stimulate sales (such as Amazon's "suggest" feature), or may assist with forecasts in the very short run. Computers excel at estimating such numerical correlations; however, they are less capable of assessing visual patterns. Data visualization therefore leverages the human capability to discover such patterns, thereby stimulating further scrutiny of data [37]. As indicated by the preceding ideas, a conducive environment must be created within which data can be managed, visualized, and analyzed [65].

Given an analysis which integrates discovery and creativity through direct observation of the data, human intuition, and strong model development, a communicative challenge remains when working with big data. Documenting and communicating relevant data sources is one of these. However, it pales in comparison with the ability to communicate the methodology whereby insights were generated. This is a challenge which has to be overcome in order to create an environment through which others, such as policy makers or other collaborators, can make sense of the results and understand how to use them. The idea of an "Analytics Cloud Environment" discussed above indirectly addresses this by giving users the opportunity to explore how various model assumptions affect results. As potential task complexity grows, institutions will have to face the challenge of considering the cost of analysis using big data as well as integrating feedback into the use of big data in order to adjust its use [65].

## Insights and their implementation

Addressing challenges with respect to the data

itself--its generation, features, and interpretation--is insufficient, as elucidated in section two, to interact meaningfully with the data. The interpretation must effectively give rise to insights which can then be acted upon to effect the desired change in the world. At one level, the very kinds of insights of interest can present their own challenges or considerations. At another, the process through which insights are translated into action must be strengthened by overcoming several intertwined challenges. Each of these aspects is considered, in turn, below.

### *Insight scope and purpose*

As discussed in section two, one way to think about the kinds of insights to generate is in terms of scope. Big data may be used better to understand the situation of a microregion, while another may be used for an entire continent, or even the whole world. Clearly, the larger the scope, the greater the potential data required. More subtle, however, is the fact that the broader the scope, the greater the potential diversity of datasets, especially when no singly-managed dataset covers the entire scope of interest. Analysts will need to ensure that datasets can interact with each other to form an accurate representation of the whole. This requires overcoming data management challenges as well as making careful decisions about aggregation to ensure helpful comparisons among the various datasets.

Alternatively, insights of interest may be considered in terms of their primary purpose. Is the use of big data meant to generate insights which heighten awareness, deepen understanding, and/or enhance forecasting? If the primary purpose is to enhance awareness, then it is likely that capabilities around visualization will be especially important, since datasets will have to be shown in an accessible way to enable shared human understanding. If, instead, the primary purpose is to understand or forecast, capabilities around data modeling will be of primary importance. This is especially the case for endeavors which seek to understand systems or processes. By utilizing sophisticated inductive techniques such as machine learning, for example, forecasting may be improved through the use of additional variables and functional forms while not necessarily enhancing understanding

of the causal web underlying a system's functioning. Understanding a system requires more than showing statistical significance; it requires the development of sensible, internally consistent models of behavior.

### *Insight implementation process*

Given a sufficient degree of clarity on the kinds of insights of interest, some thought should be given to the ways in which the analysis of data will be translated into actions which will change those original behaviors of interest. As discussed in section two, such an implementation process requires defining next steps, creating structures to take them, and, finally, actually taking those steps. Below are discussed some considerations and challenges which are connected with all three aspects.

### *Project vs. process*

Whether big data is used for a project or process will impact the degree to which a landscape analysis should be conducted to understand what data sources were used in the past to accomplish similar objectives. In addition, if an institution seeks ongoingly to utilize big data to assist in managerial tasks, then such a process-orientation will need to place emphasis on technical capacities to analyze as well as institutional capacities to act upon and ongoingly inform such analysis.

### *Culture*

Challenges at the level of culture can have a substantial bearing on the fruitful execution of a project endeavor which utilizes big data. At one level, institutions may avoid releasing data due to paternalistic beliefs that others cannot make use of the data or due to concerns that the data will be used against them [33]. Overcoming cultural challenges such as these will be particularly critical when opening data for citizen engagement [15].

Beyond a culture which stifles data openness, several beliefs or perspectives may inhibit the use of big data by institutions. One study showed that public administrators often have three major viewpoints about big data. Some view big data as presenting an opportunity to improve services and solve practical tasks. Others view big data as a fad which will not necessarily improve services. Instead, these think that big data may actually strengthen governmental control of the masses. Finally, others believe that, while big data may offer opportunities for the government to reach more citizens and tailor services, use of big data will not necessarily constitute an actual improvement. In fact, those individuals who are not electronically connected may be marginalized [21]. Particularly in contrast to the last two views, those in the public sector seeking to realize the value of big data will have to face the challenge of creating a culture which seeks to improve systems and processes based on data<sup>3</sup> [5].

Even if a single institution is convinced of the benefits of using big data, a culture of cooperation and collaboration among multiple institutions is critical [28]. At one level, cooperation is critical for the establishment of standards by which various agencies or individuals may access and use data. Beyond this, collaboration would help elucidate the possible uses of big data and help nurture and channel energies to utilize it. One example of this is the Global Partnership on Development Data recommended by a panel in the United Nations concerned with poverty. This Partnership would bring "together diverse but interested stakeholders" who would, as a first step, "develop a global strategy to fill critical gaps, expand data accessibility, and galvanize international efforts to ensure a baseline for post-2015 [poverty] targets" [21]. Whether in public or non-profit sectors, cooperation among managers and elected politicians is essential to use big data to inform decision making [12]. One way to begin collaborations with an institution as a whole is to find those individuals within it who are willing to share data and collaborate

---

3. One author has noted that, where such a data-driven culture exists, the tendency of conceiving of society as simply a set of sub-populations rather than as a single social body will have to be avoided. Although this appears to be primarily a philosophical concern, it may well have real implications on defining the public good and defining the means to achieve it, including the role of government [13].

on generating insights for policy or management decisions [26].

### *Use and abuse*

Another set of challenges that pertain to big data relate to people's interactions with the results or prescriptions from big data analyses. In some cases, a challenge that must be overcome is the possibility that many organizations, companies, and institutions would rather avoid the use of big data since it may reveal truths which they would prefer remain hidden. Such non-truth-seeking behavior could include pharmaceutical companies avoiding the use of big data to assist with understanding the effects of a drug once introduced in the market. Alternatively, healthcare companies may avoid health personalization via big data since it will encourage patients to use more preventive medicine thus reducing their income from doctor visits [37].

As the potential for big data is increasingly understood, the potential for it to be used for purposes contrary to the public good and/or ethical principles is just beginning to be explored. Such uses of big data often leverage the seeming predictability of human behavior. Some cases have already begun tentatively to demonstrate the ability to use big data, for example, to infer when an individual is dating [45], the strength of a marriage [46], or the effect of the size of a retail markdown sign on sales [39]. Such predictive information may be used for various purposes, not all of which are malicious. However, the fact that big data permits such predictions raises concerns about ensuring that it is used in a way which harmonizes with the public good.

At one level, such powerfully predictive deductions from big data raise the challenge of ensuring that people are aware of what companies or institutions

know about them and the way that data is utilized [37]. Beyond awareness, people need to be educated on the value of their data and how to control its dissemination through means as simple as, for example, privacy settings on social media [45]. At an even higher level, substantive ethical questions need to be discussed in the public sphere regarding the use of predictive, albeit somewhat imperfect, information by companies or other institutions seeking to advance their own interests [39].

Due to the somewhat unstoppable nature of the use of big data, dealing with such challenges at an institutional level will likely have to be handled through management, rather than preventing adoption [39] [37], such as by increasing transparency on how data is used [54]. Institutions may have to move beyond hierarchical, centralized, and rule-driven structures in order to deal with the variety of uses and abuses of big data [37].

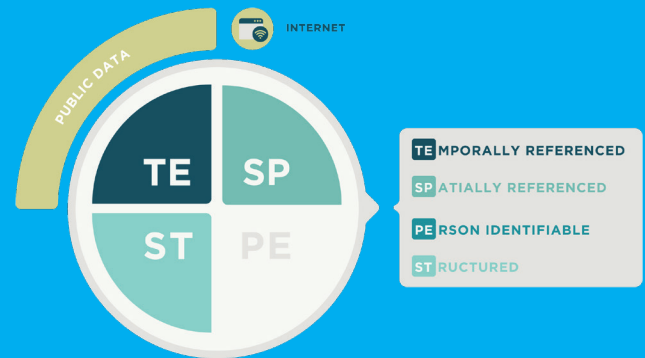
Another, more subtle challenge arises when the use of big data is integrated into systems and outputs. When big data is systematically incorporated to generate some kind of output, such as predictions on crime hot spots or product preferences, the very people whose behavior or preferences are being predicted may consciously interact with the system in a manipulative way. One example of this is "google bombing," or cases in which people craft webpage features to trick the Google search algorithm in order to raise their page-rankings. Another dangerous possibility is for criminals to outsmart police agencies relying on big data. In particular, criminals could use the same data and algorithms to infer the police's expectations thereby enabling them to behave in unexpected ways [37]. These are examples of a broader class of problems arising from the study of social systems as compared to less self-aware systems. The subject may, in fact, become the object of its own study.

# CASE STUDY

## Using Google Trends to nowcast economic activity in Colombia

### DATA PALETTES

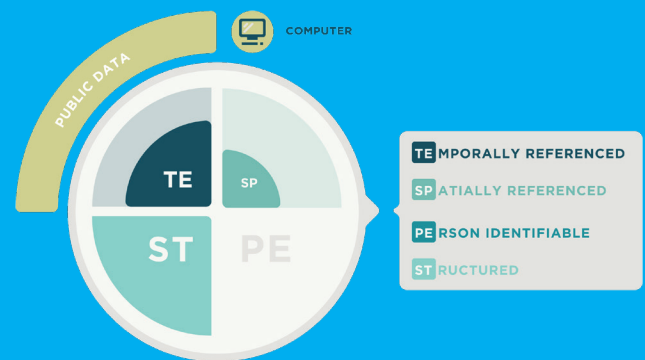
Google Trend Data



### Motivation

For the well-timed design of economic policy, it is desirable to count with reliable statistics that allow constant monitoring of economic activity by sectors, preferably in real time. However, it is a well-recognized problem that policymakers must make decisions before all data about the current economic environment is available. In Colombia, the publication of the leading economic indicators that the Administrative Department for National Statistics (DANE- for its acronym in Spanish) uses to analyze economic activity at the sectorial level has an average lag of 10 weeks. In this framework, the Ministry of Finance in Colombia looked for coincident indicators that allow tracking the short-term trends of economic activity. In particular, the approach was to exploit the information provided by the Google search statistical tool known as “Google Trends”.

GDP Official Data



### Data Generation

The data for this study comes from Google web searches. Based on the web searches performed by Google users, Google Trends (GT) provides daily information of the query volume for a search term in a given geographic region (for Colombia, GT data are available at the departmental level and also for the largest municipalities). Each individual Google Trend series is relative and not an absolute measure of search volume. That is, the period in which the search interest of a keyword is highest within the dates of inquiry receives a value of 100. All other periods for an individual series are measured relative to this highest period. To assess the performance of the indexes built using GT data, official economic activity data (both at the aggregate level and at the sectorial level) from the DANE are used. Both GT and DANE data are publicly available.

### Data Interpretation

In order to exploit the information provided by GT data, it is critical to choose a set of keywords that can be used as a proxy of consumer behavior or beliefs. In some sense, GT data takes the place of traditional

consumer-sentiment surveys. For example, the use of data for a certain keyword (such as the brand for a certain product) might be justified in the case a drop or a surge in the web searches for that keyword could be linked to a fall or an increase in its demand and, therefore, a lower or higher production for the specific sector producing that product. The analysis carried out by the Ministry of Finance in Colombia identifies keywords meaningfully related to the different economic sectors and leverages GT data for these keywords to produce leading indicators for economic activity at the sectorial level (ISAAC – for its acronym in Spanish). It is important to highlight that this approach is used only for some of the key sectors of the economy (such as agriculture, industry, commerce, construction, and transports). The performance of other sectors (such as mining or financial services or personal services) cannot be assessed using web searches and other leading indicators need to be used. Once the sectorial ISAACs are produced, the information is used to produce an aggregate leading indicator for the economic activity in the country (ISAAC+).

## Insight Implementation

The research carried out by the Ministry of Finance in Colombia showed the potential of web searches information to explain economic activity variations for some specific sectors in Colombia (in particular, agriculture, industry, commerce, construction, and transports). GT queries allow the construction of leading indicators which determine in real time the short-term trend of the different economic sectors, as well as their turning points. In addition, the leading indicator produced for aggregate economic activity (ISAAC+) shows a high correlation with its reference variable (quarterly GDP annual variation) capturing its turning points and short-term behavior. The production of these leading indicators reduces the lag associated with the publication of traditional statistics and help policy makers in the country to make timely decisions. The main limitation of this work is that the level of Internet penetration in Colombia is still relatively low (about 50%) and this implies that GT data reflects information from just a part of the country's consumers: those who have access to the Internet and use Google for their web searches. As Internet penetration deepens in the future, the representativeness of GT data will improve and make the ISAAC indicators even more relevant.

## From Idea to Ongoing Process

The research project, led by Luis Fernando Mejía, General Director of the Macroeconomic Policy Department in the Ministry of Finance, raised interest inside and outside Colombia. The ISAAC indicators produced with GT data are currently published on a monthly basis in reports which show the historical correlation between the ISAAC and GDP data at the sectorial level and highlight sectorial trends projected by the ISAAC indicators. Other countries are looking at this interesting project and might start producing similar big data driven forecasts in the future.

## References

L. F. Mejía, D. Monsalve, Y. Parra, S. Pulido and Á. M. Reyes, "Indicadores ISAAC: Siguiendo la actividad sectorial a partir de Google Trend," Ministerio de Hacienda y Crédito Público, Bogotá, 2013. Available: [http://www.minhacienda.gov.co/portal/page/portal/HomeMinhacienda/politicaFiscal/reportesmacroeconomicos/NotasFiscales/22 Siguiendo la actividad sectorial a partir de Google Trends.pdf](http://www.minhacienda.gov.co/portal/page/portal/HomeMinhacienda/politicaFiscal/reportesmacroeconomicos/NotasFiscales/22%20Siguiendo%20la%20actividad%20sectorial%20a%20partir%20de%20Google%20Trends.pdf). [Accessed 28 August 2014]

Correspondence with Luis Fernando Mejía, General Director of Macroeconomic Policy, Ministry of Finance and Public Credit, Republic of Colombia



## SECTION 6

# RETROSPECT AND PROSPECT

With the advent of computing, humanity has entered a new era of an unprecedented and exponentially rising capacity to generate, store, process, and interact with data. Seeking ways to maximize efficiency and increase product offerings by leveraging such data, the private sector has gained substantial experience over the last few decades. Such experience undoubtedly built upon scientific capabilities to analyze data and practical business acumen in order to ensure effective, real-world applications. In the public sphere, leveraging various, often very large, data sources to effect real improvements has only just begun in the last decade. The cases in this document testify to the promise for such data to be used to enhance perception, deepen understanding, and hone forecasting abilities.

Although experience with big data is relatively nascent, several conclusions can already be drawn. For data to be effective, it must be seen in the context of an ongoing process to better understand and interact with the world. In this light, use of big data should begin with a question and a description of the behaviors of interest. Use of big data from various sources requires effective and scalable methods to access, manage and interpret data. Data must be well-documented to ensure traceability. Models through which data is interpreted must be carefully selected to correspond to the data generating process. Collaboration among practitioners, social scientists, and data scientists will be critical in order to ensure the requisite understanding of the real-world conditions, data generation mechanisms, and methods of interpretation are effectively combined. Such collaboration will also enable the overcoming of major technological, scientific, and institutional challenges. Technical challenges include managing

unstructured data, linking diverse data sets, and scaling systems to respond to increasing volumes of data. Scientific challenges include the rising demand for data scientists, developing appropriate models to work with large and diverse data sets, dealing with selection bias in various forms, and communicating analytical results in order to yield actionable insights. Institutional challenges include limited access to data, cultures that don't promote learning, ethical concerns about the utilization of personal data, and the lack of standards with regard to data storage.

As the technological capacities to generate, store, and process data continue unabated, humanity will need to develop a corresponding measure of various technical, social, cultural, and institutional capabilities to ensure that big data is used toward helpful and effective ends such as strengthening early warning for disasters, enhancing awareness by providing real-time feedback, or better understanding social systems. The necessary capabilities enable the integration of big data into ongoing processes rather than one-time projects, thereby enabling its value to be continually released and refined. Spaces will be needed in which such technical, cultural, and institutional capabilities can commensurately develop. For example, members of various institutions, corporations and governments may convene to develop a shared perspective on the usefulness of big data for poverty reduction and agree to standards on its utilization. Given the variety and pervasiveness of the necessary capabilities to utilize big data to address big problems, collaborative spaces are needed to enhance the capacity of individuals, organizations, businesses and institutions to elucidate challenges and solutions in an interactive manner, strengthening a global culture of learning to reduce poverty and promote shared prosperity.





# REFERENCES

- [1] H. Varian, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, vol. 28, no. 2, p. 3–28, 2014.
- [2] Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025), 60-65.
- [3] V. Mayer-Schonberger and K. Cukier, *Big data: A Revolution that will transform how we live, work and think.*, New York: Houghton Mifflin Harcourt Publishing Company, 2013.
- [4] I. T. Union, "The World in 2013: ICT Facts and Figures.," [Online]. Available: <http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx>. [Accessed January 2014].
- [5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity.," McKinsey & Company, 2011.
- [6] D. Fisher, R. DeLine, M. Czerwinski and S. Drucker, "Interactions with big data analytics," *Interactions*, vol. 19, no. 3, pp. 50-59, 2012.
- [7] R. Bucklin and S. Gupta, "Commercial use of UPC scanner data: Industry and academic perspectives," *Marketing Science*, vol. 18, no. 3, pp. 247-273, 1999.
- [8] F. Diebold, "A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline," Working Paper, 2012.
- [9] S. Lohr, "The age of big data.," *New York Times*, 2012.
- [10] United Nations Global Pulse, "Big Data for Development: Challenges and Opportunities.," United Nations, 2012.
- [11] A. McAfee and E. Brynjolfsson, "Big data: the management revolution," *Harvard business review*, vol. 90, no. 10, pp. 60-68, 2012.
- [12] M. Milakovich, "Anticipatory Government: Integrating big data for Smaller Government," in Oxford Internet Institute "Internet, Politics, Policy 2012" Conference, Oxford, 2012.
- [13] D. Boyd and K. Crawford, "Six provocations for big data," in *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.
- [14] The World Bank, "Open Data Essentials," 2014. [Online]. Available: <http://data.worldbank.org/about/open-government-data-toolkit/knowledge-repository>. [Accessed July 2014].
- [15] L. Hoffmann, "Data mining meets city hall," *Communications of the ACM*, vol. 55, no. 6, pp. 19-21, 2012.
- [16] J. Manyika, M. Chui, D. Farrell, S. V. Kuiken, P. Groves and E. A. Doshi, "Open Data: Unlocking Innovation and Performance with Liquid Information.," McKinsey & Company, 2013.

- [17] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins and N. Kruschwitz, "Big data, analytics and the path from insights to value.," MIT Sloan Management Review, vol. 21, 2014.
- [18] S. Lucas, Interviewee, [Interview]. December 2013.
- [19] A. Asquer, "The Governance of big data: Perspectives and Issues," in First International Conference on Public Policy, Grenoble, 2013.
- [20] Centre for Economics and Business Research, "Data equity: unlocking the value of big data.," SAS, 2012.
- [21] United Nations, "A New Global Partnership: Eradicate Poverty and Transform Economies Through Sustainable Development.," United Nations Publications, New York, 2013.
- [22] J. Blumenstock, Interviewee, [Interview]. January 2014.
- [23] V. Frias-Martinez, Interviewee, [Interview]. December 2013.
- [24] M. Khouja, Interviewee, [Interview]. December 2013.
- [25] A. Siegel, Interviewee, [Interview]. December 2013.
- [26] N. Eagle, Interviewee, [Interview]. December 2013.
- [27] A. Leshinsky, Interviewee, [Interview]. January 2014.
- [28] R. Kirkpatrick, Interviewee, [Interview]. February 2014.
- [29] A. Cavallo, Interviewee, [Interview]. February 2014.
- [30] N. Scott and S. Batchelor, "Real Time Monitoring in Disasters," IDS Bulletin, vol. 44, no. 2, pp. 122-134, 2013.
- [31] S. Thornton, Interviewee, [Interview]. February 2014.
- [32] United Nations Global Pulse, "Big Data for Development: A Primer," United Nations Publications, 2013.
- [33] W. Hall, N. Shadbolt, T. Tiropanis, K. O'Hara and T. Davies, "2012," Nominet Trust, Open data and charities.
- [34] A. R. Syed, K. Gillela and C. Venugopal, "The Future Revolution on Big Data," Future, vol. 2, no. 6, 2013.
- [35] E. Bertino, P. Bernstein, D. Agrawal, S. Davidson, U. Dayal, M. Franklin, ... and J. Widom, "Challenges and Opportunities with Big Data," White Paper, 2011.
- [36] World Economic Forum, "Big Data, Big Impact: New Possibilities for International Development," World Economic Forum, Geneva, 2012.
- [37] D. Bollier, "The Promise and Peril of Big Data," The Aspen Institute, Washington, D.C., 2010.
- [38] M. Ghasemali, Interviewee, [Interview]. January 2014.
- [39] R. Cross, Interviewee, [Interview]. December 2013.
- [40] E. Wetter and L. Bengtsson, Interviewees, [Interview]. January 2014.
- [41] K. Warner, T. Afifi, T. Rawe, C. Smith and A. De Sherbinin, "Where the rain falls: Climate change, food and livelihood security, and migration," 2012.
- [42] M. Hilbert, "Big Data for Development: From Information-to Knowledge Societies," Working Paper, 2013.
- [43] L. F. Mejía, D. Monsalve, Y. Parra, S. Pulido and Á. M. Reyes, "Indicadores ISAAC: Siguiendo la actividad sectorial a partir de Google Trend," Ministerio de Hacienda y Crédito Público, Bogotá, 2013. Available: <http://www.minhacienda.gov.co/portal/page/portal/HomeMinhacienda/politicafiscal/>

reportesmacroeconomicos/NotasFiscales/22  
Siguiendo la actividad sectorial a partir de Google  
Trends.pdf. [Accessed 28 August 2014]

[44] L. T. Interviewee, [Interview]. December 2013.

[45] R. Vasa, Interviewee, [Interview]. January 2014.

[46] R. Smolan and J. Erwit, *The human face of big data.*, Sausalito, CA: Against all odds productions, 2012.

[47] N. Eagle, M. Macy and R. Claxton, "Network Diversity and Economic Development," *Science*, vol. 328, no. 5981, pp. 1029-1031, 2010.

[48] IBM Global Business Services, "Analytics: The real-world use of big data in financial services," IBM Global Services, Somers, NY, 2013.

[49] NYSE Euronext, "NYXdata > Data Products > NYSE Euronext > FlexTrade," 2013. [Online]. Available: <http://www.nyxdata.com/nysedata/Default.aspx?tabid=1171>. [Accessed February 2014].

[50] L. Barrington, Interviewee, [Interview]. January 2014.

[51] PriceStats, "PriceStats," 2014. [Online]. Available: <http://www.pricestats.com>. [Accessed February 2014].

[52] A. Robertson and S. Olson, "Sensing and Shaping Emerging Conflicts: Report of a Joint Workshop of the National Academy of Engineering and the United States Institute of Peace: Roundtable on Technology, Science, and Peacebuilding," The National Academies Press, 2013.

[53] UNICEF, Regional Office for Central and Eastern Europe and the Commonwealth of Independent States, "Tracking Anti-Vaccination Sentiment in Eastern European Social Media Networks," 2013.

[54] A. Howard, *Data for the public good*, O'Reilly Media, 2012.

[55] OECD, "New Data for Understanding the Human Condition: International Perspectives," OECD, 2013.

[56] M. Lagi, K. Bertrand and Y. Bar-Yam, "The food crises and political instability in North Africa and the Middle East," arXiv:1108.2455., 2011.

[57] S. Greengard, "Policing the Future," *Communications of the ACM*, vol. 55, no. 3, pp. 19-21, 2012.

[58] IBM Institute for Business Value, "Analytics: The real-world use of big data.," 2012.

[59] B. Wescott, Interviewee, [Interview]. January 2014.

[60] V. E. M. Lehdonvirta, "Converting the virtual economy into development potential: knowledge map of the virtual economy," *InfoDev/World Bank White Paper*, pp. 5-17, 2011.

[61] K. Kitner, Interviewee, [Interview]. December 2013.

[62] S. Lohr, "Sure, Big Data is Great. But So Is Intuition," *New York Times*, 29 December 2012.

[63] SecondMuse, "Random Hacks of Kindness 2013 Report," 2013.

[64] United Nations Global Pulse, "Mining Indonesian Tweets to Understand Food Price Crises," United Nations Publications, 2014.

[65] F. Almeida and C. Calistru, "The main challenges and issues of big data management," *International Journal of Research Studies in Computing*, vol. 2, no. 1, 2012.

[66] B. Fecher, "Open Science: One Term, Five Schools of Thought," in *The 1st International Conference on Internet Science*, Brussels, 2013.

## ANNEX 1

# SELECTED BIBLIOGRAPHY

### Big data, general topics

- Almeida, F. L. F., & Calistru, C. (2012). The main challenges and issues of big data management. *International Journal of Research Studies in Computing*, 2(1).
- Asquer, A. (2013). The Governance of Big Data: Perspectives and Issues. Retrieved from: <http://ssrn.com/abstract=2272608> or <http://dx.doi.org/10.2139/ssrn.2272608>
- Bertino, E., Bernstein, P., Agrawal, D., Davidson, S., Dayal, U., Franklin, M., ... & Widom, J. (2011). Challenges and Opportunities with Big Data. Whitepaper presented for Computing Community Consortium.
- Bollier, D., & Firestone, C. M. (2010). The promise and peril of big data (p. 56). Washington, DC, USA: Aspen Institute, Communications and Society Program.
- Boyd, D. & K. Crawford. (2011). presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011.
- Bucklin, R. E., & Gupta, S. (1999). Commercial use of UPC scanner data: Industry and academic perspectives. *Marketing Science*, 18(3), 247-273.
- Centre for Economics and Business Research. (2012). Data equity: unlocking the value of big data. Centre for Economics and Business Research White Paper, 4, 7-26.
- Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. *Interactions*, 19(3), 50-59.
- Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025), 60-65.
- IBM Institute for Business Value. (2012). Analytics: The real-world use of big data. Executive Report. Retrieved from <http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html>
- LaValle, Steve, et al. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review* 52.2 : 21-31.
- Lohr, S. (2012). Sure, Big Data is Great. But So is Intuition. Retrieved from [http://www.nytimes.com/2012/12/30/technology/big-data-is-great-but-dont-forget-intuition.html?\\_r=0](http://www.nytimes.com/2012/12/30/technology/big-data-is-great-but-dont-forget-intuition.html?_r=0)
- Lohr, S. (2012). The age of big data. *New York Times*. Retrieved from <http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60-66.

Smolan, R., & Erwitte, J. (2012). *The human face of big data*. Sausalito, CA: Against all odds productions.

Syed, A. R., Gillela, K., & Venugopal, C. (2013). The Future Revolution on Big Data. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(6), 2446-2451.

## Big data case studies

Eagle, N., Macy, M., and Claxton, R. (2010). Network Diversity and Economic Development. *Science*, 328(5981): 1029-1031. Retrieved from: [http://realitymining.com/pdfs/Eagle\\_Science10.pdf](http://realitymining.com/pdfs/Eagle_Science10.pdf)

Greengard, S. (2012). Policing the Future. *Communications of the ACM*, 55(3), 19-21

Herrera, J. C., Work, D. B., Herring, R., Ban, X. J., Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C*, 18, 568-583.

Hoffmann, L. (2012). Data mining meets city hall. *Communications of the ACM*, 55(6), 19-21.

Lagi, M., Bertrand, K. Z., & Bar-Yam, Y. (2011). The Food Crises and Political Instability in North Africa and the Middle East. arXiv preprint arXiv:1108.2455.

Milakovich, M. (2012). Anticipatory Government: Integrating Big Data for Smaller Government, paper presented at the Oxford Internet Institute "Internet, Politics, Policy 2012" Conference, Oxford, 20-21 September.

Robertson, A. & Olson, S. (2013). Sensing and Shaping Emerging Conflicts. *The National Academies Press*. 13-14. Retrieved from: [http://www.nap.edu/catalog.php?record\\_id=18349](http://www.nap.edu/catalog.php?record_id=18349)

Robertson, C., Sawford, K., Daniel, S. L., Nelson, T. A., & Stephen, C. (2010). Mobile phone-based infectious disease surveillance system, Sri Lanka. *Emerging infectious diseases*, 16(10), 1524.

Scott, N., & Batchelor, S. (2013). Real Time Monitoring in Disasters. *IDS Bulletin*, 44(2), 122-134.

UNICEF, Regional Office for Central and Eastern Europe and the Commonwealth of Independent States. (2013). Tracking Anti-Vaccination Sentiment in Eastern European Social Media Networks. Retrieved from: <http://www.unicef.org/ceecis/Tracking-anti-vaccination-sentiment-in-Eastern-European-social-media-networks.pdf>

## Big data for development

Hilbert, M. (2013). Big Data for Development: From Information-to Knowledge Societies. Available at SSRN 2205145.

Howard, A. (2012). *Data for the public good*. O'Reilly. Karlsrud, J. (2014). Peacekeeping 4.0: Harnessing the Potential of Big Data, Social Media, and Cyber Technologies. In *Cyberspace and International Relations* (pp. 141-160). Springer Berlin Heidelberg.

Lehdonvirta, V., & Ernkvist, M. (2011). Converting the virtual economy into development potential: knowledge map of the virtual economy. *InfoDev/World Bank White Paper*, 1, 5-17.

Organisation for Economic Co-operation and Development. (2013). *New Data for Understanding the Human Condition*. Retrieved from <http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.htm>

United Nations Global Pulse. (2012). *Big Data for Development: Challenges & Opportunities*. UN, New York: NY

United Nations Global Pulse. (2013). *Big Data for Development: A Primer*. Retrieved from <http://www.unglobalpulse.org/bigdataprimer>

World Economic Forum. (2012). Big Data, Big Impact: New Possibilities for International Development. Retrieved from <http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development>

## Open data / Open science

Fecher, B., & Friesike, S. (2013). Open Science: One Term, Five Schools of Thought (No. 218). German Council for Social and Economic Data (RatSWD).

Hall, W., Shadbolt, N., Tiropanis, T., O'Hara, K., & Davies, T. (2012). Open data and charities. Nominet Trust. Retrieved from <http://www.nominettrust.org.uk/knowledge-centre/articles/open-data-and-charities>

McKinsey Global Institute. (2013). Open Data: Unlocking innovation and performance with liquid information. Retrieved from: [http://www.mckinsey.com/insights/business\\_technology/open\\_data\\_unlocking\\_innovation\\_and\\_performance\\_with\\_liquid\\_information](http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information)

## ANNEX 2

# INTERVIEW LIST

### **Aaron Siegel**

Head of Interaction and Online Experience, Fabrica Benetton

Topic areas: data visualization

### **Alberto Cavallo**

Billion Price Project Lead

MIT – Sloan School of Management

Topic area: economics, price indices

### **Anthony Leshinsky**

Media Services Lead

Coldlight Solutions

Topic area: business intelligence, data analytics

### **Bill Wescott**

Executive Vice President

The CoSMo Company

Topic areas: satellite data, systems modeling

### **Bayan Bruss, Mohamad Khouja, Jian Khoddad, Jon Pouya Ehsani**

Co-Founders

Logawi

Topic areas: Text analytics and analysis

### **Erik Wetter**

Assistant Professor, Stockholm School of Economics

Co-founder, Flowminder

Topic areas: mobile data, big data for development

### **Graham Dodge**

CEO and Founder

Sickweather

Topic areas: Social media data

### **Jean Francois Barsoum**

Senior Managing Consultant, Smarter Cities Water and Transportation, IBM

Topic areas: big data, urban infrastructure and operations

### **Joshua Blumenstock**

Assistant Professor

University of Washington

Topic areas: mobile and social media data

### **Kathi Kitner**

Senior Researcher, Anthropologist

Intel Labs

Topic area: big data for development

### **Laura Crow**

Principal Product Manager

M-PESA

Topic areas: mobile data and financial data

### **Linus Bengtsson**

Executive Director

Flowminder/Karolinska Institutet

Topic areas: mobile data, big data for development

### **Luke Barrington**

Senior Manager, Research and Development

Digital Globe

Topic areas: satellite data, big data analysis

### **Mahyar Ghasemali**

Partner and Co-founder

dbSeer

Topic areas: data infrastructure and processing

### **Mohamad Khouja**

Data Scientist and Big Data Solutions Specialist

IREX/Logawi

Topic areas: opinion mining, sentiment and lexical analysis

### **Nathan Eagle**

CEO

Jana

Topic areas: mobile data, big data for development

### **Rajesh Vasa**

Senior Lecturer, Faculty of Information and Communication Technologies

Swinburne University of Technology

Topic areas: social media big data

### **Renato de Gusmao Cerqueria**

Senior Manager, Natural Resources Solutions

IBM Research Brazil

Topic areas: data analytics and urban systems

### **Robert Kirkpatrick**

Director

UN Global Pulse

Topic areas: big data for development

### **Robin Cross**

Research Director

DemandLink

Topic areas: Retail data and predictive analysis

### **Sean Thornton**

Research Fellow

Data-Smart City Solutions

Topic areas: data and public policy

### **Shannon Lucas**

Senior Enterprise Innovation Manager

Vodafone

Topic areas: mobile and financial data

### **Vanessa Frias-Martinez**

Research Scientist

Telefonica Research

Topic area: mobile data



## ANNEX 3

# GLOSSARY

**3 "V"s** - A term defining certain properties of big data as volume (the quantity of data), velocity (the speed at which data is processed) and variety (the various types of data).

**Algorithm** - A formula or step-by-step procedure for solving a problem.

**Anonymization** - The process of removing specific identifiers (often personal information) from a dataset.

**API (Application Programming Interface)** - A set of tools and protocols for building software applications that specify how software components interact.

**Business intelligence** - The use of software tools to gain insight and understanding into a company's operations.

**Clickstream analytics (analysis)** - The collection, analysis and reporting of data about the quantity and succession of mouse clicks made by website visitors.

**Crowdsourced** - The collection of data through contributions from a large number of individuals.

**Data cleaning/cleansing** - The detection and removal, or correction, of inaccurate records in a dataset.

**Data exhaust** - Data that is collected as a digital by-product of other behaviors.

**Data governance** - The process of handling and management of data being utilized in an endeavor, including policies, data quality and risk management.

**Data migration** - The transition of data from one format or system to another.

**Data science** - The gleaning of knowledge from data as a discipline that includes elements of programming, mathematics, modeling, engineering and visualization.

**Data silos** - Fixed or isolated data repositories that do not interact dynamically with other systems.

**Data warehousing** - the practice of copying data from operational systems into secondary, offline databases.

**Geospatial analysis** - a form of data visualization that overlays data on maps to facilitate better understanding of the data.

**Hadoop** - an open source platform for developing distributed, data-intensive applications.

**Internet of things** - The unique digital identifiers in objects that can automatically share data and be represented in a virtual environment.

**Latency** - the delay in the delivery of data from one point to another, or when one system responds to another.

**Machine learning** - The creation of systems that can learn or improve themselves on the basis of data; often linked to artificial intelligence.

**Map/reduce** - a method of breaking up a complex problem into many chunks, distributing them across many computers and then reassembling them into a single answer.

**Mashup** - The use of data from more than one source to generate new insight.

**Metadata** - Information about, and descriptions of, data.

**Nowcasting** - A combination of "now" and "forecasting," used in both meteorology and economics referring to immediate term forecasting on the basis of real time data flow.

**Open data** - Public, freely available data.

**Open science** - An initiative to make scientific data and research openly accessible.

**Petabyte** - 1 thousand terabytes.

**Predictive analytics/modeling** - The analysis of contemporary and historic trends using data and modeling to predict future occurrences.

**Quantitative data analysis** - the use of complex mathematical or statistical modeling to explain, or predict, financial and business behavior.

**Reality mining** - The study and analysis of human interactions and behavior through the usage of mobile phones, GPS and other machine-sensed environmental data.

**Relational database** - A database in which information is formally described and organized in tables representing relations.

**Sentiment analysis (Opinion Mining)** - The use of text analysis and natural language processing to assess the attitudes of a speaker or author, or a group.

**Structured data** - Data arranged in an organized data model, like a spreadsheet or relational database.

**Terabyte** - 1 thousand gigabytes.

**Text analytics** - the process of deriving insight from unstructured, text-based data.

**Topic modelling** - The use of statistical models or algorithms to decipher themes and structure in datasets.

**Tweet** - a post via the Twitter social networking site restricted to a string up to 140 characters

**Unstructured data** - Data that cannot be stored in a relational database and can be more challenging to analyze from documents and tweets to photos and videos.

**Visualization** - Graphic ways of presenting data that help people to make sense of huge amounts of information.

Photo credits: cover page: Mor Naaman / flickr.com; page 2 Martin Sojka / flickr.com; page 7: Uncultured / flickr.com; page 12: CGIAR Climate / flickr.com; page 25: CGIAR Climate / flickr.com; page 35: Linh Nguyen; page 42 / unsplash.com : NASA's Marshall Space Flight Center / flickr.com; page 54 / unsplash.com; page 56: Wojtek Witkowski / flickr.com





**THE WORLD BANK**

IBRD • IDA | WORLD BANK GROUP

Latin America & the Caribbean



**WORLD BANK GROUP**

Macroeconomics & Fiscal Management

SECONDMUSE 