Data Science for Global Health

Peter Lance, John Spencer, Mark Janko

October 2016

USAID
FROM THE AMERICAN PEOPLE

MEASURE Evaluation

Data Science for Global Health

Peter Lance, John Spencer, Mark Janko

October 2016

Cover image: Adapted from "The Great Wave off Kanagawa," a woodblock print by Hokusai (circa 1830–1833)
ISBN: 978-1-943364-99-2

MEASURE Evaluation
University of North Carolina at Chapel Hill
400 Meadowmont Village Circle, 3rd Floor
Chapel Hill, NC 27517 USA
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. TR-16-143

ABBREVIATIONS

GIS	geographic information system
LMIC	low- and middle-income country
USAID	United States Agency for International Development

CONTENTS

INTRODUCTION

Ours is an age of explosive growth in data. Traditional data sources are ever deeper and richer with each passing day. Even more rapidly, new sources of powerful data are emerging. The result is a stunning, exponentially growing torrent of data from every corner of the globe and about nearly every dimension of human life and activity.

This offers challenges and opportunities for society in general and for global health professionals in particular. The increasing amount of data can lead to more insight, better policy and programs, and improvements in people's lives. However, data can also create noise and confusion if it isn't used effectively.

This whitepaper introduces global health professionals to data science. Data science is a production process for generating actionable information. It helps us find, understand, and communicate knowledge hidden in the growing data deluge. In global health, successful data science efforts can extract value from data that might otherwise go unused, and use it to inform policy and support programmatic decision making.

The data environment for global health is undergoing rapid changes. Take, for instance, the evolution of regional, national, and local routine information systems addressing health. The quality and reliability of these systems have improved dramatically in recent years and in more and more areas of the globe. Beyond these systems, the growth in survey capacity in organizations around the world has led to an increase in the range and frequency of survey activity. Moreover, the trends in information capture through surveys and routine systems familiar to many in public health, universities, government, and the public sector have been accompanied by advances in the private sector that are equally if not more stunning.

A prehistoric "selfie" from Algeria indicates that 10 millennia ago, the Sahara was savannah, not desert. Source: Gruban; Creative Commons

These developments extend to the dawn of the human story. For instance, the rock paintings of the Paleolithic and Neolithic periods are among the earliest known examples of self-generated data, revealing across an ocean of time how our ancestors saw themselves and the world around them. With the development of agriculture and the rise of increasingly sophisticated societies, we see the first attempts at purposeful data collection. Examples are the 2 C.E. Han Chinese censuses and the seminal 1086 Domesday Book, which essentially surveyed every person, parcel of land, and physical asset in England. So the explosion in the availability of data that we are witnessing now is part of a process that has been building for a long time.

And yet, in scope, depth, and pace of evolution, we have never before seen anything like what is happening globally right now. The power and sophistication of traditional data capture systems in global public health: for example, censuses; routine information systems; sentinel surveillance systems; and ongoing routine survey programs, such as the Demographic and Health Surveys funded by the United States Agency for International Development (USAID) is increasing. At the same time, newer types of data are rapidly becoming available: for example, social media data; routine public and private enterprise data; and routine capture of data from the devices that drive our lives, which is the basis for our so-called "digital exhaust" (Sarasohn-Kahn, 2014). The confluence of these two trends will allow us to learn things about our world that we could not until now. For instance, in the past 15 years, we have witnessed the emergence of Internet navigation tools (Google, Yahoo, Bing, etc.), social media tools (Facebook, Weibo, Twitter, Instagram, Snapchat, etc.) and online communities that are providing information about how we see ourselves and the world around us in a fashion directly

analogous to the way the cave paintings once did. The differences now are the volume, speed, cost, and detail of the information.

At the center of this revolution are the rapidly falling costs of data transmission, capture, and storage, which have made it possible to observe and organize humanity's thinking about the world in ways that were impossible even 15 years ago. The pace of information expansion from all of these sources shows no sign of slowing. Indeed, the smart money at this point is on breathtaking acceleration of these developments.

Many metaphors for this shift come to mind, but the one we'll use here is the *Data Tsunami*. The challenge is how to extract actionable information from the Data Tsunami in the most deliberate, rapid, accurate, and impactful fashion so that it can serve to improve health and other facets of human welfare.

Data science is a discipline and practice devoted to recovering the actionable information from the data in the tsunami. Data science begins with the identification of need: What sort of information would fill gaps in effective health policymaking? Answering this question involves not only recognizing the existence of gaps in the evidence base but also understanding precisely what would constitute "actionable information" with respect to the gaps. The answer also involves understanding the subtle linkages among finding and analyzing data and communicating the findings to facilitate action. There is a symbiotic relationship between these tasks. For instance, analyzing data may prompt a search for additional data, which in turn will need analysis, and which can then be transformed into actionable data products. While these fundamental considerations are present in traditional data analysis, data science takes advantage of special skills to create a formal, purposeful production process for efficient use of the data.

One must appreciate the possibilities within the Data Tsunami. Because the tsunami is constantly growing and is at any given moment essentially unknowable in any comprehensive, detailed sense, skillful and efficient exploration is important. Moreover, one must appreciate what is required to extract data from the tsunami. Because of differences in scale, scope, or other factors, data will often need to be restructured and manipulated—a process known as *data wrangling*. The inputs for data processing and analysis—the raw data from the tsunami—might not immediately suggest the kinds of actionable information they will yield. Credible analysis often requires working with data whose collective relevance becomes clear only after skillful combination of data sources creates a platform for analysis far more useful than any of the constituent data would be on their own.

The actionable information will typically be conveyed in the form of data products. A data product is a device that distills the data from the tsunami into actionable information. Sometimes a data product is just a number, such as—for instance—an estimate of the relationship between two variables critical to the theory of change driving the program's design. However, the possibilities for data products are evolving rapidly and limited only by the imagination. Suppose, for example, that rather than supplying just a number summarizing the relationship between those two variables, one provided an application that allowed decision makers to simulate the interplay of the variables under different circumstances, and to do so in real time (on their personal computers, tablets, smartphones, etc.) as they discussed

program design. Such products transform what had been a slow, tedious, iterative process for analysts and decision makers (as the analysts continuously updated their technical advice in the face of an evolving policy discussion) into a far more dynamic, efficient, and effective one.

The final element of data science is communicating data products to decision makers. This involves insuring that the products reach the right people, who in turn clearly understand the possibilities and limitations.

In short, data science is a production process, and its central challenge is to integrate the functions just described. Much as the practice of data science often involves merging disparate data sources into a whole far more powerful than the sum of its parts, the process of data science intrinsically involves coordinating the identification of information needs, data exploration, analysis, and the communication of data products so that these activities are far more productive and effective than they could be on their own.

THE DATA TSUNAMI'S IMPLICATIONS FOR GLOBAL HEALTH

We live in an age of boundless possibility in terms of perhaps our most important resource: information. In this section, we discuss what this means, as a launching point for our vision of data science for global health.

A Global Phenomenon

The data availability revolution we are currently experiencing is unprecedented. Simply put, more data, in several senses, is becoming available from more places than ever before. This is not limited to wealthy industrial and postindustrial Western economies. In virtually every society, the amount of available data is growing. There has never been a moment in human history when we have had access to so much data, and hence so much information, about so many facets of so many societies. As the amount of data expands, the value of data as an asset, and the value forgone by not using that asset, grows at a dizzying rate.

Thus far, discussion around evidence-based decision making in global public health has tended to conceptualize data in terms of traditional sources: censuses, routine information sources (such as routine health information systems), and population surveys. And indeed, the increasing quantity of data available from these sources is part of the Data Tsunami. However, the Data Tsunami in low- and middle-income countries (LMICs) mirrors that in wealthier societies, in that new types of data are also part of it. These new types of data can answer questions that cannot be addressed (or addressed easily) by traditional data sources. Moreover, novel combinations of data from new and traditional sources can leverage traditional data in previously impossible ways, enriching the evidence that can be drawn from them. This will force us to expand our thinking about what data means in global public health—or else risk doing far less to advance the welfare of the poor than we might have done.

Consider these examples of recent data science achievements:

- Identifying rural impoverished areas using an automated process that classifies rooftops of structures using Google Earth imagery (Abelson, 2014)
- Using cell phone data to track population mobility and model Ebola transmission (Brown, 2015)
- Analyzing social media channels to reveal public attitudes toward vaccination (United Nations Children's Fund, 2013)

As important as the Data Tsunami has already been in LMICs in terms of actual data yielded, perhaps its greatest importance is the rapid expansion it has provoked of the infrastructure for generating new data in these societies. As the Data Tsunami gathers momentum, data begets data.

This is reflected in several convergent developments. Generally improving educational opportunities have increased the supply of workers capable of playing a role in capturing information. Decades of specific, targeted capacity building by multilateral and bilateral development organizations in partnership with local governments and civil society institutions have led to tremendous improvement in LMICs' ability to operate routine health information and surveillance systems, conduct surveys, and carry out other national data-capture efforts, such as large-scale socioeconomic and demographic surveys (e.g., SUSESNAS, in Indonesia) and national censuses.

In increasing numbers, the citizens of LMICs are embracing technological opportunities that the citizens of wealthier societies—for whom these tools now seem so natural and inevitable—embraced themselves only within the past decade. This has been particularly apparent in the area of "self-generated" personal data, which is becoming the most immediate and important aspect of the information revolution for many people. Later we will discuss self-generated data in more detail.

The Data Tsunami is not just happening in the rich, "post-industrial" economies that have traditionally been thought of as the primary focus for earlier phases of the information revolution (such as the personal computer, or PC, revolution). While data expansion (as well as expansion in data capture capacity) in LMICs may lag behind the rate of change in wealthier societies, in absolute terms, the increase in data richness there has been vast. And the general trend of data growth in LMICs matches what is happening in rich, postindustrial societies. This process will play out over years, not generations or decades.

So What Exactly Is in the Data Tsunami?

The Data Tsunami has many manifestations, but in the realm of global public health, a few big ones stand out.

We begin with the traditional information sources that inform health and human welfare policy: national censuses, routine reporting systems, and household surveys. The information flows from these three traditional conduits are not new, but the breadth and richness of the information emerging from them now and for the foreseeable future are unprecedented.

National Censuses

There is a long history of collecting information about population circumstances. The Babylonians conducted the first ones, more than 5,000 years ago. The oldest extant census dates to 2 C.E, during China's Han Dynasty. The Domesday Book, of 1086, collected in England and Wales at the direction of William the Conqueror, is one of the earliest European censuses.

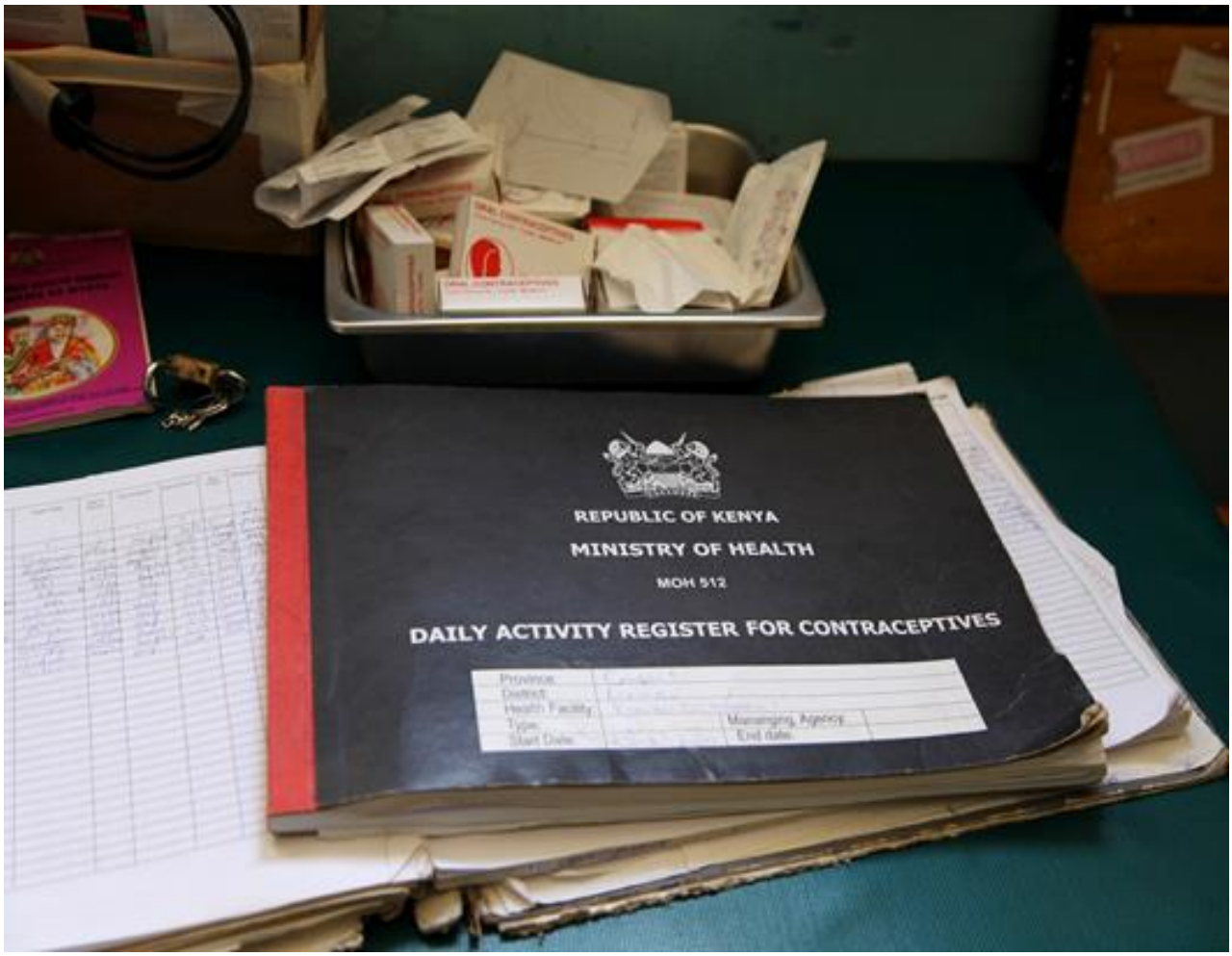
Despite their long pedigree, censuses have an uneven track record as information sources. Undertaking a census is an expensive and logistically challenging operation with several potential weaknesses. First, only very recently have many nations conducted censuses with any regularity. Second, many censuses (across poor, middle-income, and high-income societies) don't contain a great deal of information beyond that required to develop demographic profiles of the society in which they were taken. For instance, very few modern censuses match the richness of information that the Domesday Book captured.[1] Third, the quality of censuses varies, and it is not always easy to gauge a census's accuracy. Often, the poor quality of a census is not known until another data collection effort, such as a national household survey, presents discrepancies or contradictory findings. Fourth, some countries are so slow to release data that, by the time a census enters the public domain and is disseminated to government agencies, it is more useful as a historical record than as a policy tool. Fifth, often the information is not digitized or indexed in some easily linkable fashion (such as with geographic identifiers to facilitate use in a geographic information system [GIS]). This lack of an index makes it difficult to find census data for specific locations and also to link the data with other data sets. To a certain degree these last three weaknesses are an indication of limited country technical capacity to conduct censuses.

Globally, constraints caused by poor technical capacity are disappearing. Over the past decade, sustained capacity building and technical exchange (by donors, census bureaus in wealthier nations, and others), a rapidly growing information infrastructure, and rising educational levels have combined to yield a tremendous overall increase in the speed, sophistication, and reliability of censuses in countries where previous censuses have been unreliable.

Routine Reporting Systems

A routine reporting system is one designed to capture information regarding health and health system performance on an ongoing, regular basis, by means of an established continuously operating collection mechanism. Routine health reporting systems are often thought to be routine reporting of patient flows, but they frequently include financial records and reporting, operational records, and detailed patient information. This information goes beyond diagnoses made and treatments offered, to enable critical perspectives on such questions as the degree of access to care. The 2015 Ebola outbreak in West Africa highlighted the importance of routine systems as a key component of epidemiological surveillance.

[1] The Open Domesday website (http://opendomesday.org/) reveals the staggering amount of information collected in the Domesday Book in an approachable fashion.



The data tsunami isn't only electronic information. It also includes paper records, such as the ones stored in this registry in a Kenyan health facility. Photo: Wayne Hoover, MEASURE Evaluation

As with censuses, routine reporting systems are not without their faults. Information flows from routine systems have been spotty and unreliable in some countries. Additionally, over-reporting can be an issue when a single person visiting different health points is counted as multiple individuals.

As with censuses, improvement in the quality of routine reporting systems in LMICs is likely to accelerate in response to intense donor focus and the confluence of better trained staff, greater information technology possibilities, and greater urgency to improve systems at all levels. The power and capability of health systems is also likely to grow as the systems are integrated with other, parallel official mechanisms currently becoming more common, such as national identification systems.

As performance improves, these systems will generate mini-data tsunamis of their own, enabling LMICs to know more than ever before about the health and healthcare of their populations, and about the inner workings of their health systems. This will create stunning new opportunities for understanding the strengths and weaknesses of these societies in terms of health and the performance of their health systems. This can open the door to effective, evidence based policies that will allow LMICs to chart a course to full development on the health front.

Surveys

Surveys differ from censuses and routine systems, in that they seek to learn about population circumstances from samples taken from those populations—not necessarily on a routine basis. At present, surveys in LMICs essentially serve two roles.

First, they seek information not easily obtained from the other two traditional sources. For example, population-level patterns may not be identifiable from census data, because the indicators behind them are too elaborate to collect. Nor can these patterns be gleaned from routine systems, which offer profiles mainly of those who engage with the health system,[2] data geared to narrowly targeted information objectives (such as impact evaluations), and ad hoc data on special topics. Even in wealthy nations, with technically strong and reliable censuses and routine systems, information on specific topics is often obtained only through surveys. For instance, the U.S. Census gleans much of what it learns about the characteristics of the U.S. population from such surveys as the American Community Survey—not the census. In other words, surveys will remain a vital tool even with full, complete, and robust censuses and routine systems in place.

Second, surveys are often used to plug gaps in the information provided by routine systems. Surveys conducted in Bangladesh in 2001, 2010, as well as an upcoming survey to estimate maternal mortality illustrate their growing role. Surveys sampling representative groups of a population have distinct advantages over censuses or routine reporting systems, because—with their smaller samples—they are easier to conduct due to their ability to collect data by sampling representative groups of a population.

With censuses and routine systems, improvements have been achieved by addressing gaps in technical capacity or other limitations. In contrast, surveys can build on a history of success. Beginning in the 1980s, systematic survey programs, such as the World Bank Living Standards Measurement Surveys or USAID's Demographic and Health Surveys, have achieved staggering momentum with hundreds of surveys conducted to international standards using global state-of-the-art methods in survey craft. At the same time, there have been many more focused, complex, ongoing survey agendas. Examples are surveys by the University of North Carolina at Chapel Hill's Carolina Population Center (in the Philippines, China, Russia, Thailand, and elsewhere), by the Massachusetts Institute of Technology's Abdul Jamal Lateef Poverty Action Lab, by MEASURE Evaluation, and by the Health Communication Partners project and the Family Life surveys linked to the Rand Corporation and the University of California at Los Angeles (Indonesia, Malaysia, etc.).

[2] As routine systems are integrated in such mechanisms as national identification systems, they may begin to offer some of the population-level insights now yielded only by surveys.

One legacy of these survey "institutions" is an enormous advance in in-country technical capacity. Another is an expansion in the capacity of many in-country institutions, sometimes in conjunction with international partners, as a dividend of their pursuit of broad and ambitious survey agendas of their own. Finally, the past decade has seen a steady uptick in the presence of traditional commercial survey mechanisms.

Interviewers conducting a household survey. Photo: MEASURE Evaluation

As a result of this activity over the past few decades, many LMICs have deep capacity for population surveys—capacity that actually exceeds what wealthier nations had at the outset of this era, leading to a global convergence in terms of technical capacity to conduct surveys more quickly, at higher quality, and at lower relative cost than ever before.

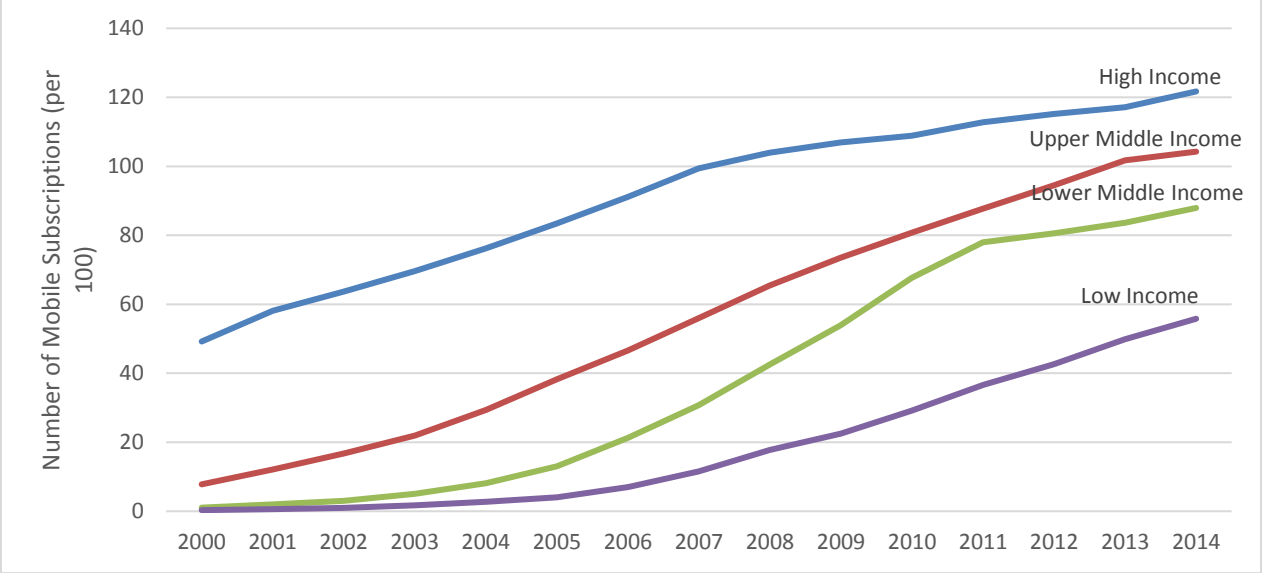
These societies are now poised to benefit from the population survey culture, which has revolutionized life in wealthier societies with the insights gleaned from the data generated.

New Sources of Data

As exciting as all of this is, traditional mechanisms are just the leading edge of the Data Tsunami. In every sense, the information revolution is arriving in nations throughout the world.

Consider just two manifestations of this: the mobile revolution and the explosion in internet connectivity. In just the past decade, mobile phones have become commonplace consumer products nearly everywhere. Figure 1 shows trends in mobile subscriptions by countries classified by World Bank Income Group. Across every group, the trajectory is upwards.

Figure 1. Mobile subscription trends by World Bank income groups



Source: International Telecommunication Union, World Telecommunication/ICT Development Report and database, retrieved from http://data.worldbank.org/indicator/IT.CEL.SETS.P2?cid=GPD_31&end=2014&locations=XD-XM-XN-XT&start=2000

Moreover, revolutions are already occurring *within* the mobile revolution. For example, even in the poorest parts of Africa, the traditional "feature phones"[3] market is collapsing as consumers rapidly shift to smart phones. Indeed, by 2019 "feature phones" will account for less than a third of the mobile phone market in Africa and the Middle East (McLeod, 2015). By the first quarter of 2015, half of the mobile phones sold in Africa were smartphones (such as those based on Google's Android or Apple's iOS); sales of such phones have been growing by nearly 70 percent a year (McLeod, 2015). As fast as Africans adopted mobile technology, now they are upgrading to smart phones at an even faster rate.

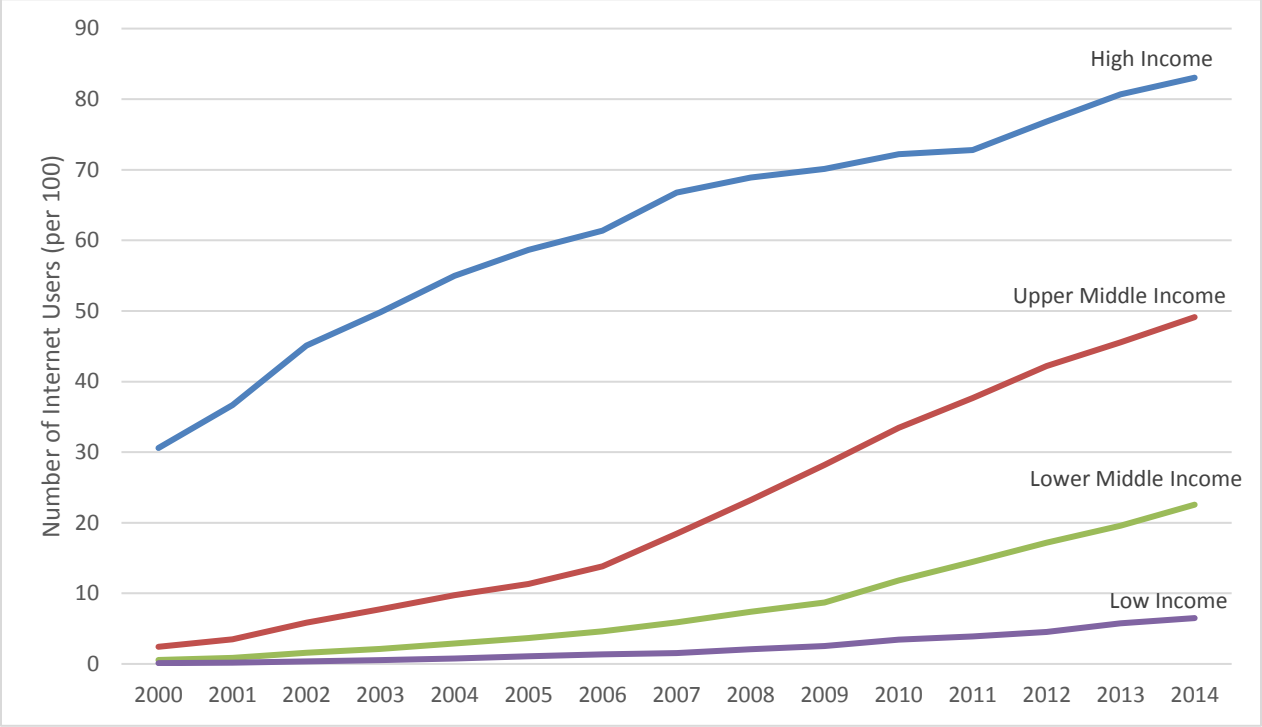
[3] "Feature phone" is a term used to describe a phone with only basic features. Typically, feature phones do not allow users to connect to the Internet or download apps.

Girls in a tricycle in the Philippines. Photo: Adam Cohn
https://www.flickr.com/photos/adamcohn/15348525562/in/photolist-poicV5

The progress has been just as dramatic for Internet penetration. Figure 2 shows the sharp increase in the number of people acquiring access to the Internet over the first 15 years of the century, regardless of country income group.

Example of a feature phone.
Photo: Clive Darra
https://www.flickr.com/photos/osde-info/680364521/

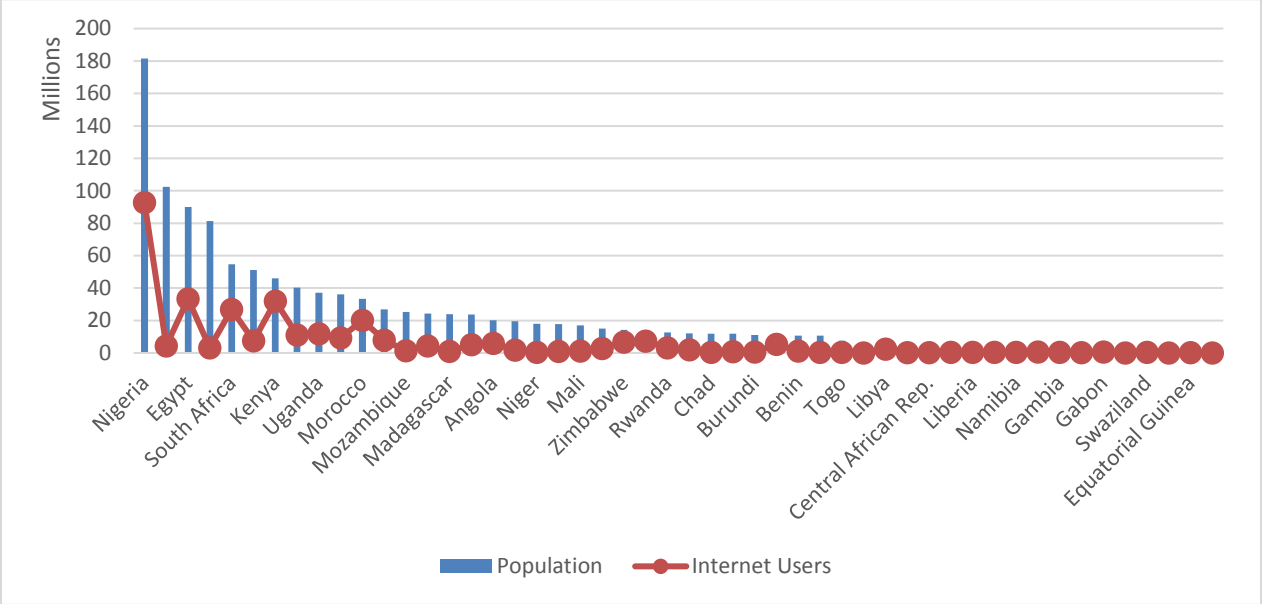
Figure 2. Number of Internet users per 100 by World Bank income groups



Source: International Telecommunication Union, World Telecommunication/ICT Development Report and database, and World Bank estimates. Retrieved from http://data.worldbank.org/indicator/IT.NET.USER.P2?cid=GPD_31&end=2014&locations=XD-XM-XN-XT&start=2000

Figure 3 shows the adoption of Internet use in Africa. In 2016, Africa had just over 330 million Internet users (Internet World Stats, 2015). We can see from the figure that four countries—Nigeria, Egypt, Kenya, and South Africa—are the continent's focal points for this revolution, serving as regional diffusion points for a spreading Internet culture in Africa.

Figure 3. Population and Internet users in Africa, as of June 2016 (millions)



Source: internetworldstats.com, retrieved from
http://www.internetworldstats.com/stats1.htm retrieved 09/23/16

As fascinating as these developments around mobile phones and Internet adoption are, they are not themselves elements of the Data Tsunami. So what do they have to do with it?

Quite simply, they will enable the tsunami's continual expansion, by serving as the means by which billions of people worldwide become data generators. As people engage with the world using these information tools, they reveal vast amounts of information about themselves and their environment. In some sense, we are becoming respondents in a gigantic, fascinating, ever-shifting, detailed, noisy, chaotic, rich, brilliant, global survey (on present trends soon to be essentially a census), in which each of us defines the survey questions with our online behavior.

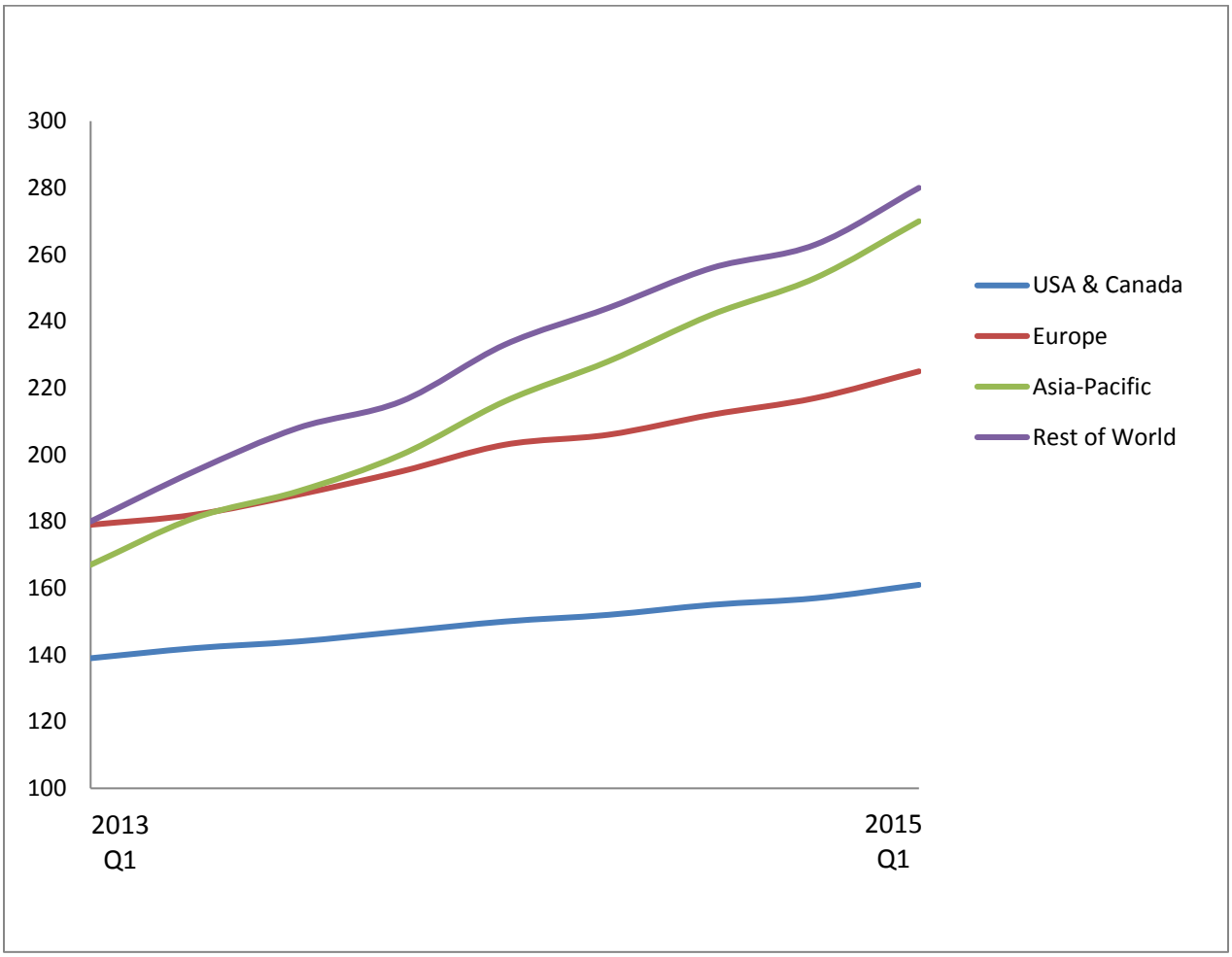
This growth in access to the Internet and the capacity that goes with it constitute an ever-expanding set of conduits through which the information that people generate can flow. Consider social media, which did not exist (at least in any currently recognizable form) 20 years ago. Today, the list of social media tools—prominent among them Facebook, Twitter, Weibo, Instagram, Snap Chat, Tumblr, Pinterest, Google+, YouTube, and blogging tools such as WordPress—is growing at a dizzying pace.

Figure 4 offers evidence of the developing world's embrace of social media. In fact, the market for one of the most iconic social media products, Facebook, is now essentially flat in North America, its place of origin. All of the action in terms of growing market share is in the Asia-Pacific region and, crucially, the rest of the world.

In part, the flat market for Facebook in North America reflects the maturity of its presence there, as of the time frame considered in the figure. Its story is probably typical of that of other social media

products: After becoming widely popular in wealthier countries, Facebook then quickly diffused globally. At present, the United States is more or less the locus of social media development, with most of the remainder in Europe and developed Asia (Japan, the Republic of Korea, and Taiwan primarily). Weibo, in China, is one of a few local exceptions, but it has not caught on globally. That said, as the deepening culture of social media combines with the growing armies of computer programmers, and as information technology entrepreneurs and venture capitalists come together in places such as India, we are likely to see LMICs become exporters of social media products globally.

Figure 4. Facebook users, 2013–2015 (millions)



Source: http://www.internetworldstats.com/facebook.htm, retrieved July 2015

Figure 5. A world lit by Twitter

This screenshot maps all tweets around the world for a few hours on June 28, 2015. Source: Tweetping.net

For all of its facets and ever-evolving character, the global social media revolution has one clear implication: it is turning everyone into a data producer. Yes, much of that data is simply noise: it is unlikely that social media posts such as "American Idol So Crazy #Can'tSingForAnything" or "I'm bored #need2getalife"—to pick at random two recently trending hashtags—will meaningfully inform decisions in public health or any other policy arena. But if one looks beyond the frivolous, the real implication of social media is that they are turning billions of people into real-time reporters of their immediate circumstances and experiences in ways that could inform policy decisions. Moreover, this reporting is essentially instantaneous—a data collection mechanism much faster than formal data collection mechanisms could be.

Consider sentinel epidemic modeling. This typically requires reporting by healthcare providers, which lags the actual outbreak of disease. Social media can potentially offer essentially real-time crowd-sourced sentinel monitoring. As an outbreak occurs, people at its epicenter may use social media to report symptoms that they and those around them are experiencing. Social media can be monitored to capture deviations from normal patterns that might portend an outbreak. Critical information about the locus of the outbreak, illness course, and social responses can be quickly aggregated, providing a nearly immediate profile of the outbreak. Of course, an outstanding and constantly

evolving challenge will be to develop analytical tools that most effectively separate the wheat from the chaff of social media chatter.

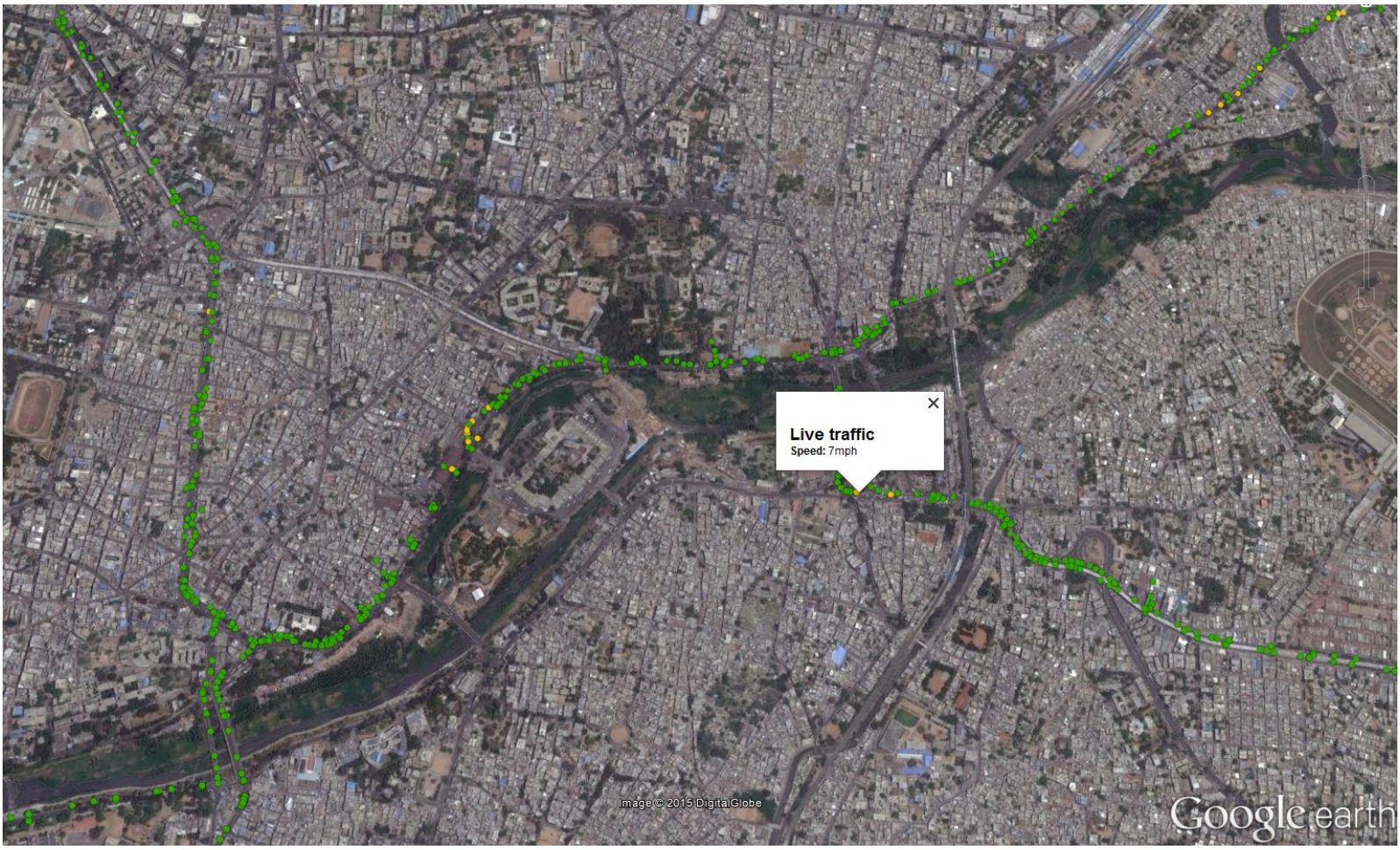
This information will not replace traditional surveillance mechanisms, but it can augment them tremendously. Failing to exploit such information will lead to blind spots in policy decisions.

Is it the sniffles or the start of something big?

Another example of self-generated data is the location information automatically reported by certain smart phone apps. This type of data can be used to report traffic and mobility patterns. Until recently, such information was limited to wealthy societies. This is changing, and rapidly. Figure 6 shows Google traffic pattern data (of the sort readily available for New York, Boston, London, Paris,

Moscow, and virtually all small- and medium-sized cities in rich countries) for Hyderabad, India. The image was taken from the standard, publicly available release of Google Earth.

Figure 6. Real-time traffic in Hyderabad, India



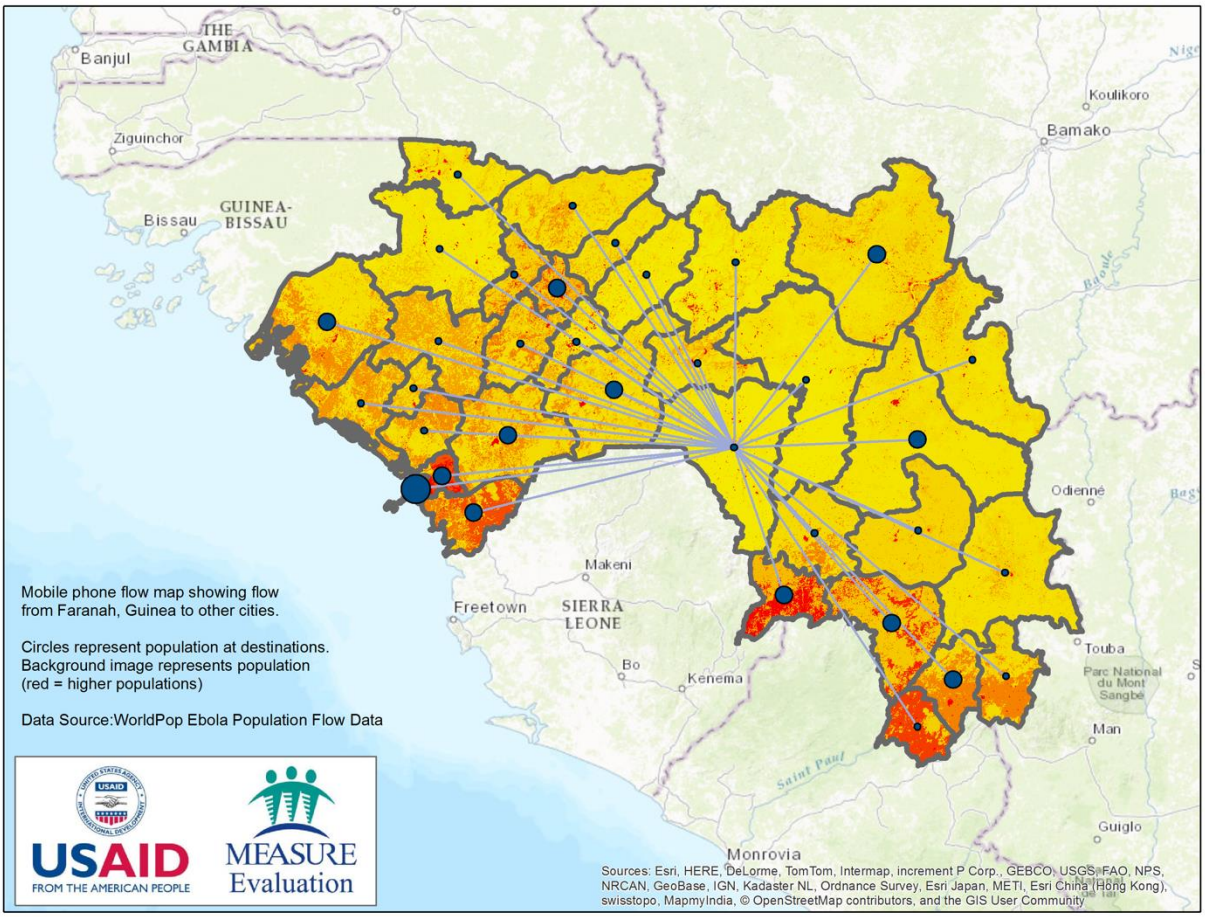
This screenshot showing live traffic in Hyderabad, India, was retrieved on July 28, 2015. Source: Google Earth

Google Traffic and similar products illustrate the possibilities latent in data that smart device users create—not only active data (e.g., texts, Facebook posts, phone calls, etc.) but also passive data (tower pings, automatic server pings from apps, etc.). The potential applications of this passive data are enormous. Consider the challenge of sampling for surveys in dynamic, densely populated cities where the urban fringe is constantly expanding in a process driven primarily by migrants from rural areas who settle in emergent, often informal slum communities. These types of population dynamics are often missed by censuses, which tend to be widely spaced in time. Typically, data on mobile phone use can be difficult or costly to obtain, but they can reveal much about population dynamics. For instance, during the 2015 Ebola crisis in West Africa, mobile phone data were made available

showing population flows within countries (Worldpop, n.d.). This information has helped epidemiologists understand migration patterns and identify areas of risk.

In 2015, Faranah, Guinea experienced an outbreak of Ebola. Mobile phone data make it possible to see where people typically travel. For example, Figure 7 shows considerable outmigration from Faranah to heavily populated areas around Conakry—the capital—as well as other cities in southern Guinea. Information like this can help epidemiologists anticipate the spread of a contagious disease such as Ebola.

Figure 7. Population flows from Faranah, Guinea



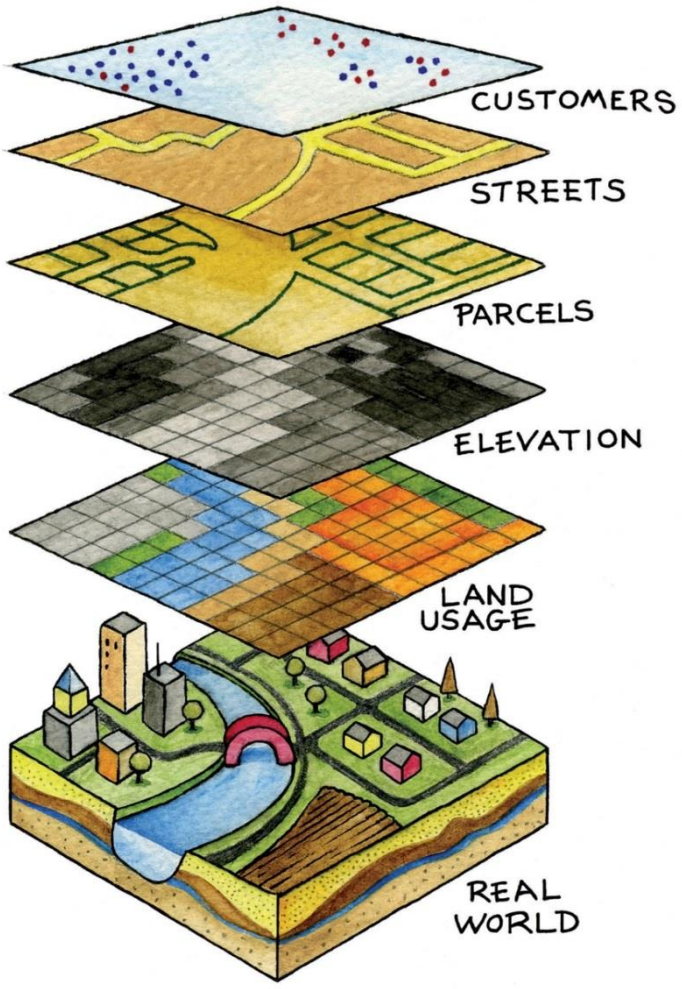
Mobile phone flow map showing population flow from Faranah, Guinea to other Guinean cities. The size of the circles represent the number of people traveling from Faranah to the destination. The shading of the polygons represent the population of the entire prefecture. Data source: Worldpop. Map: MEASURE Evaluation

Surveys to obtain this knowledge would be costly and slow. Routine reporting systems aren't designed to capture patients' travel behavior. The Data Tsunami makes the timely discovery of such patterns possible.

The Promise of Geographic Information Systems

The Data Tsunami's greatest value can be realized when novel and disparate streams of information are combined. The basic idea is that the whole of combined data is more powerful than the sum of its parts. This idea is well-established in GIS applications.

Figure 8. GIS: The vintage view



GIS is an approach to synthesizing information through a common geography. In laymen's terms, you merge information based on the spatial units (administrative units, user-created polygons, etc.) from which that information was generated. Thus, you might have three disparate types of information (e.g., average temperature in a given month from weather data sources, modern contraceptive use from survey data, and number of distinct cell phones from tower pings) for a country. Because these data share the common geography of being in the same location, a GIS can

link them. This possibility is often captured in the "layered" conceptualization of a GIS (see Figure 8), whereby the various types of information are represented by distinct layers placed on top of one another, aligned by their common GIS.

The result can be a rich way of looking at the world—one that already has yielded commonplace consumer products. For instance, today ordinary consumers with Internet access can tap into Google Earth, which—though hardly cutting edge in comparison with other GIS systems making full use of the data tsunami—is a vast source of information by any historical measure. This is one of a growing array of apps making increasingly sophisticated used of GIS on mobile phones.

Figure 9. The Google Earth view of Rio has layers

Each symbol on this Google Earth map of Rio de Janeiro, retrieved December 2015, represents a data element.

We are now at a point where extraordinary and sometimes new possibilities are emerging for the extraction of information using common geography. These possibilities are driven, roughly, by two factors:

- The geographical indexing of information is becoming nearly universal, often in subtle, passive ways
- The technology for efficiently extracting and merging useful information is improving rapidly

We are on the verge of an era in which powerful image analysis software will allow the algorithmic identification of faces, signs, text strings, logos, building types, gender, etc. from images. This will extend tremendously the information that can be extracted from self-generated data, such as photographs taken throughout the world and posted on social media sites.

In addition to the rise of mobile devices, the spread of social media, the increasing ubiquity of geographic information stamping, and the availability of tools to mine social media for information, the amount of routine enterprise systems data has exploded. Its source is credit card companies. The Data Tsunami allows all kinds of enterprises to record all sorts of information about their operations— often passively and without expense. This can include cost information, customer details, inventory and stock information, and virtually every other recordable aspect of operations.

The back end of a modern enterprise: the server room at the United Kingdom's National Archives. Source: Wikimedia Commons

This development first gathered momentum in the corporate world in developed nations. A desire to extract value from the mountains of information that these corporations generate gave rise to data science as a discipline. Now, however, these practices have spread throughout the globe. For instance, in the past few years, there has been a surge in the sophistication of the enterprise information systems used by the nongovernmental organizations that implement global development programs, and by the governments of some LMICs, as well. As this capacity grows, information of great interest for policy and programming that has been difficult or even prohibitively expensive to track—such as implementation cost—should be readily available.

The Impact of Data Capture on the Data Tsunami

A revolution in the means and scope of data capture has done much to make the data tsunami possible. Take, for instance, GIS. GIS exists in large part due to another technological revolution: the emergence of a robust satellite infrastructure. Sputnik, the first satellite, was launched in October, 1957. Over the next several decades, more and more countries began launching satellites, which grew in number and sophistication, and with a broadening range of purposes: communications, navigation, photography, weather tracking, and espionage. There are now roughly 2,300 satellites orbiting the earth—so many that collision is a growing concern and the debris from them is an ongoing hazard for all orbital devices.

Satellite technology accelerated the GIS revolution. It provided the georeferenced imagery that is a foundation of many geographic information systems. Satellite-based positioning systems rapidly provide a highly accurate location of practically anywhere on the planet. These advances allowed the world to be viewed in terms of a common geography, and did so with a speed, precision, and cost that were impossible in the past. The geographic information thus captured has enhanced tremendously the power and richness of GIS.

From this survey of the Data Tsunami, three things should be clear. First, more data and more kinds of data are becoming available from more places than ever before. Second, this process is as forceful in LMICs as it is in wealthy, postindustrial economies. Third, this wave will continue to grow, at speeds even faster than we have already witnessed.

How Does The Data Tsunami Relate to "Big Data"?

The information revolution is not the only hyper-transformative aspect of the modern era. Indeed, the pace of technological, social, cultural, economic, environmental, political, and institutional change is so rapid that some have suggested that history itself is accelerating. Every one of these facets of rapid change seems to bring with it buzzwords that carry intuitive or emotional weight but little concrete meaning.

Perhaps the best example of this new vocabulary is the phrase *Big Data*. We are told that Big Data is transforming our professional and personal lives. But what exactly is Big Data? What are its implications? What possibilities and challenges does it present?

Many seem to suggest that Big Data literally means enormous data sets. One definition, often cited, is data that are "high volume, greater variety and quicker velocity" (Raftree, et al., 2014). Focusing on the massiveness of available data is tempting since, to be sure, there has been a rapid increase in the number of data sets of stupefying size. But is this definition of Big Data really useful? In terms of benefits, although more observations can yield more precise estimation, they do so with a diminishing return. More observations will not by themselves necessarily remove a fundamental identification challenge, such as confounding. More variables are sometimes useful (they can mitigate confounding, for example) but many of the variables in large data sets are essentially useless for answering some particular question, and the set that is not useless is often insufficient for answering that question.

So what might serve as a useful way of thinking about Big Data? The best answer rests on a simple, central truth: Big Data *is* the Data Tsunami. They are one and the same.

The challenges of the Data Tsunami arise only in part from the tsunami's size. One could possess an infinite supply of data about every aspect of the world, but even then it would be necessary to find the relevant data, analyze them, and communicate the findings. High-volume, high-velocity data are nice to have, but their value is the knowledge one can extract from them. Even the most expansive definition of the term *Big Data* cannot convey the process needed to capture the data's value.

What Does The Data Tsunami Mean For Global Health?

This Data Tsunami represents a stunning increase in the amount of potentially actionable information that can be marshaled for better decision making. By actionable information, we mean information that can directly inform choices regarding priorities, strategic approaches, policies, programs, and allocation of scarce resources. Simply put, within the tsunami is the potential to learn far more about the world as we operate in it. Another way of looking at the tsunami, however, is that as it grows, the cost of ignoring it and making less well-informed decisions, which lead to poorer outcomes, grows ever steeper.

It was exactly this realization that first led corporations in wealthy countries to contemplate how to exploit the information in the tsunami. By ignoring that information, these corporations were making poorer decisions than they might otherwise have made, with mounting consequences for profits.

The global public health community does not enjoy the luxury of a metric of success as clear and reducible as profit, but the same basic challenge applies. By failing to make the most use of available information, we don't make the best decisions possible. And that has implications for our bottom line: more maternal and child deaths, higher fertility rates, less-educated societies, more violence against women, slower response times to emergent epidemics, and so on.

In the absence of a profit/loss consequence, the international community has come together to propose a set of metrics to measure social progress. The Sustainable Development Goals (SDGs) and the preceding Millennium Development Goals (MDGs) represent the world's effort to understand progress on a wide range of health and public welfare elements. Data are at the core of both; indeed, the drive to achieve the SDGs is known as "the data revolution" (Data Revolution Group, 2014).

So who will make the most of this great wave of information? In the simplest terms, the answer to this question will come down to the answers to four questions.

Who Will Best Understand the Information Gaps Confronting Decision Makers?

Usefully informing decision making is impossible without a comprehensive understanding of the challenges that decision makers face. This requires understanding what the decisions are, the questions on which these decisions hinge, and the evidence base on which answers to those questions rests. More subtle but just as important, it requires recognizing the vital questions that aren't being asked, either because of an incomplete understanding of the information the Data Tsunami offers or because of a failure to exploit information within the tsunami that would allow more facets of an issue to be understood.

Who Can Recognize the Emerging Data?

The tsunami is rapidly evolving, with stunning growth in size and scope. Every day, more information, and more types of information, flow from more places in the chaotic laboratory of the human experience. The vastness and complexity of the tsunami is one of the reasons many people are slow to recognize its value.

To successfully exploit the Data Tsunami, one must have some sense of what it contains. This is not an insurmountable challenge. For instance, we take search engines such as Google, Yahoo, or Bing for granted and forget the problem they solve: navigation of the big, brilliant, mess known as the World Wide Web. They do this largely through a remarkably simple process: indexing the pages of the web by key words and metadata within them, allowing for instantaneous returns from an essentially infinite number of potential search needs.

The challenge confronting us with the Data Tsunami is much the same as the challenge that gave rise to indexing, metadata, and all the rest in the context of the World Wide Web two decades ago. This challenge needs to be met by bright, creative people who will make work on it their specialty. They

will need to develop innovative tools and approaches. And they will need to understand the elegant brilliance of simplicity for these tools and approaches to be as successful as search engines.

Who Can Develop and Exploit the Most Cutting-Edge Methods for Data Synthesis and Analysis?

The data extracted from the tsunami will likely not exist in a form that immediately yields actionable information for some great decision of the day. Instead, it will need to be cleaned, merged, and refined. Then, the information within the data will need to be extracted, a process that typically uses the most powerful tools of statistics, econometrics, computer science, machine learning, and a host of related analytical fields—tools that lead to the most precise and effective actionable information. Finally, the actionable information should be packaged in a form that maximizes its value for decision makers. This means exploiting the full range of possible ways to convey the information, from the traditional report framework to dynamic, interactive information applications. The outcome of this process is a data product: the information extracted from the Data Tsunami. Raw data flow into data products and actionable information flows out in the most effective manner possible.

Who Can Convey the Insights Uncovered in the Most Effective Way?

Data products, and the actionable information they contain, are useless if the decision makers who are the intended audience do not understand them or recognize their value or relevance. Competing with the cacophony of voices and overload of information that is life in the modern world, critical information must be presented and positioned in a fashion that allows it to be seen and heard. Some data products will require explanation, demonstration, and perhaps even training for decision makers to grasp them. This process of communication of data products is an essential, strategic art.

DATA SCIENCE: DELIVERING DATA PRODUCTS OUT OF THE INFORMATION TSUNAMI

It's not necessary to overthink it, because a powerful and simple approach is right in front of us. Above all, data science is the process of producing actionable data products. Specifically, it is the process of recognizing questions that can be answered by the data tsunami, marshaling the right information from the Data Tsunami, making that information actionable, packaging it into the most useful data product possible, communicating the data product to consumers, and doing all of this in a timely fashion. This multifaceted discipline is data science.

The Data Tsunami Moves Fast

The pace at which the world generates data means that an efficient process for finding, analyzing, and using data is essential. Without that, decisions will always be made based on the way things *were*—not how they *are*. It is far better to be surfing the crest of the Data Tsunami than drifting awash behind it.

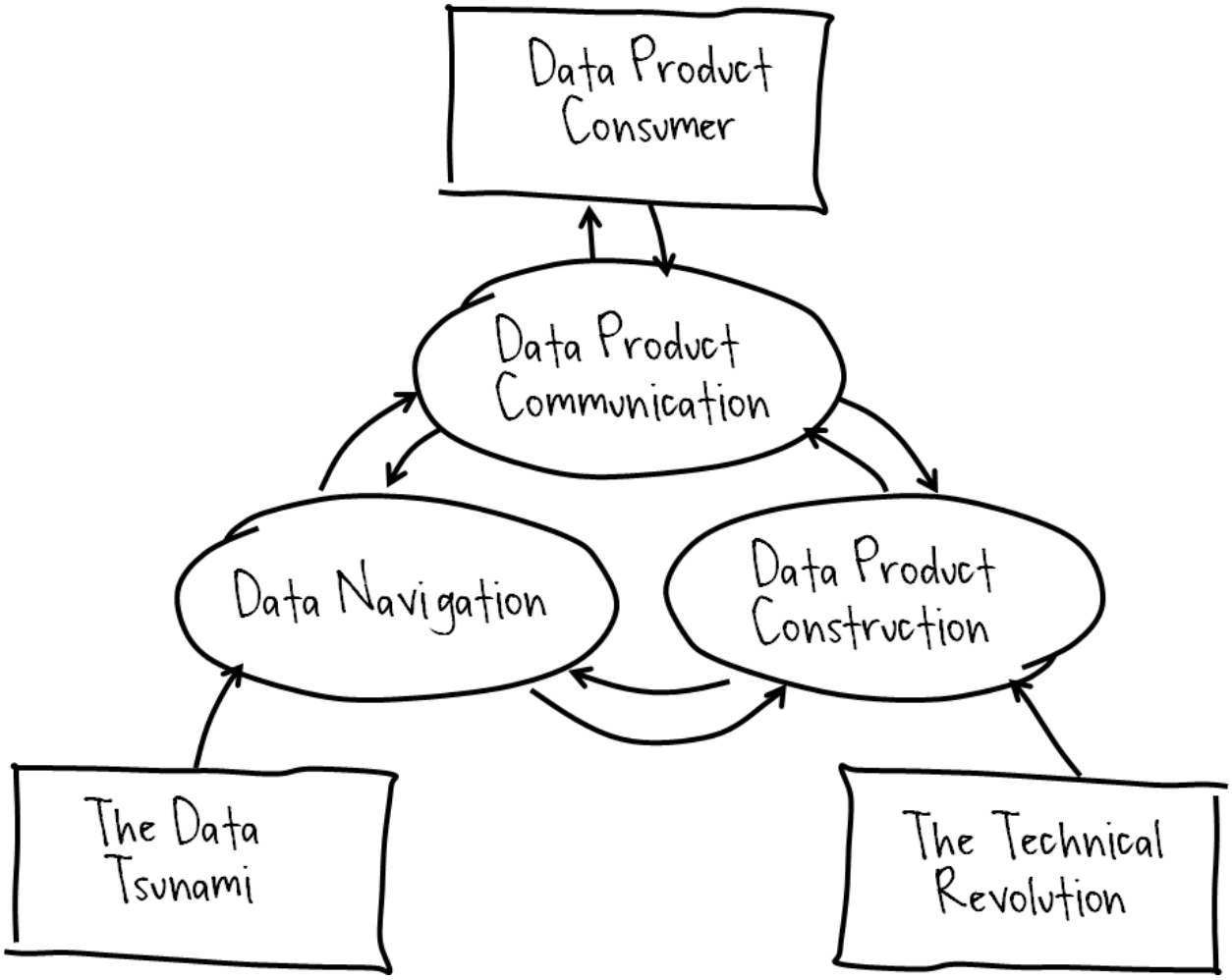
The speed and size of the Data Tsunami offer so many new opportunities that it's possible to move beyond the traditional approaches of analyzing and using data. Traditionally, global health professionals rely on data from multiple sources such as health management information systems, household surveys conducted at some point in the past, and population counts from a census already out of date (assuming the numbers were right to begin with). For data from a routine reporting system, they wait weeks or months for the data to be entered (or longer, if no electronic system exists). The data then must be pulled together, analyzed, written up, and published, adding even more months to the lag between the data and reality.

The work took place in a linear fashion: (1) wait for data to become available; then acquire it; (2) combine multiple datasets and conduct analysis around a narrow set of specific questions; and (3) produce a report or other data product. There is little opportunity for new realities on the ground to be reflected in the data, analysis, or report.

Data science is more agile, not only capturing new realities on the ground but also painting a picture of what those realities are. Its agility allows practitioners to use nontraditional data sources to find unexpected facets of the factors affecting health.

Data science involves three activities that must operate in a continuously synergistic fashion in order to create a stream of useful data products. Figure 10 captures the essence of this process. The ovals represent key activities of data science, the data product, and the data science team. The squares represent elements of the outside world with which the team must interact to create and successfully implement data products.

Figure 10. The data science model



Although we envision a data product production process in which a given data science team is simultaneously pursuing several parallel data products at various stages of development, it is perhaps more helpful here for us to consider the production cycle of just one data product. It starts with the information needs of a data product consumer. In the global health context, data product consumers are those who set policies and design and implement (or oversee the implementation of) programs designed to shape health and human welfare. This is a large set of institutions: multilateral organizations (e.g., the World Bank, the Asian Development Bank, etc.), bilateral agencies (USAID, the United Kingdom's Department for International Development, Australian Aid, etc.), large donor organizations (e.g., the Bill & Melinda Gates Foundation), national governments, nongovernmental organizations, and local partner institutions (icddr,b, in Bangladesh, is an excellent example) All play

large roles in the discussion revolving around improving human welfare, and their decisions would benefit tremendously from the information contained in actionable data products.

Another node in the data science process is data product communications, and the job of the team responsible for it is to understand the information gaps or shortfalls confronting the data product consumer. This is not something the consumer will necessarily tell the team overtly, but rather something to be gleaned from an intimate understanding of the consumer's decision-making challenges—knowledge that can arise only from a true relationship with the consumer.

Because the data product communications team interacts continuously and seamlessly with the data navigation and data product construction teams, all parties should have a good intuitive sense of what questions can and cannot be answered by data science.

This understanding comes from the continuous streams of communication with these other two nodes of the team. Each node working collaboratively to identify the most useful data sources, the most effective analytical approach and the most informative data product.

Once the communication team can present a rough idea of an information need and a data product to meet it, the data navigation and data product construction teams must work together to develop that product. The task of the data navigation team is to retrieve from the Data Tsunami the kind of raw data that might support the data product. However, they cannot know the data requirements for a given analysis strategy without seamless interaction with the data product construction node.

The task of the data product construction team is to transform the raw data from the tsunami into a data product. This work typically has three phases:

1. Data preparation: This is the transformation of the data from the tsunami—through cleaning, imputation, merging, and various other data manipulations—into a form that can be analyzed.
2. Analysis: In essence, this is the process by which the needed information is extracted from the data. It can involve statistics, machine learning, algorithmic application, neural networking, GIS analysis, or some combination of these or other tools.
3. Packaging: The information is then packaged in a form most effective to inform decision making. Traditionally, this would be a report, but it could also be a powerful, user-friendly app that allows policymakers to consider the implications of different assumptions in real time as their policy and programming discussions proceed.

Steve Jobs holding an Apple iPhone 4. One aspect of the success of a smartphone is its usability. Photo: Matthew Yohe, Wikimedia Commons

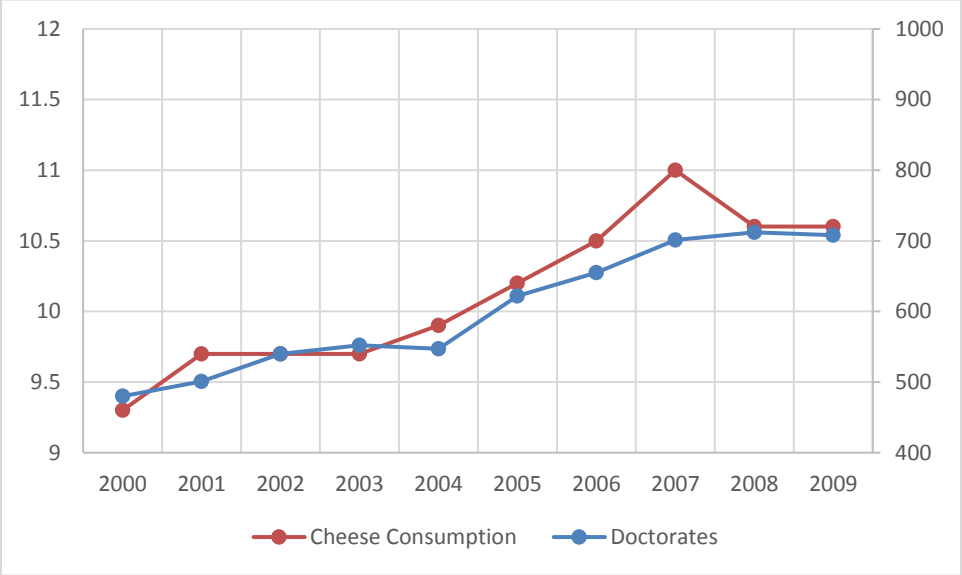
The end result of activity up to to this point is a data product that is user-friendly and useful to the work arrangements and discussion flows of the data product consumer. But this is not enough. The consumer needs to understand the product: its features, power, and limitations, and the information needs it can serve.

Data science projects don't always follow this scheme, nor is a data product always something conveyed overtly to the data product consumer. For instance, the product might instead be an ongoing analytical strategy for plugging some type of recurring information gap in the routine system.

Data navigation and data product construction involve the opportunistic but purposeful (as opposed to randomly roving) exploitation of the raw data in the Data Tsunami. It is useful to make a distinction between this work and what is often popularly described as "data mining." In essence, data mining is the attempt to find patterns in data. Although it has great usefulness, it cannot be

counted on to determine *meaningful* patterns. Just because a correlation exists does not make it meaningful, as Figure 11 illustrates.

Figure 11. Correlation of two unrelated variables: per capita consumption of mozzarella cheese and civil engineering doctorates awarded



The correlation of the two variables shown here is .958648. Source: Spurious Correlations, http://www.tylervigen.com/spurious-correlations.

In general, data mining is a poor model for describing the work of the data navigation and data product construction teams, which begins with a fairly good idea of what success will looks like (for example, a data product that closes or reduces an information gap) and hence is a more specific, purposeful, directed process.

Principles of Practice for Data Science

Data science teams should not be micromanaged or subjected to excessive process-related protocols, but that does not mean their work should not be guided by overarching principles. In this section, we discuss two particularly important aspects of this: commitment to scientific rigor and ethical research processes.

Putting the Science in Data Science

If data science *is* a science, its practice should adhere to some important and stringent standards that place it within the tradition of modern scientific inquiry. The first, and perhaps most important, of these is transparency. As a data science team works to create a data product, all of their decisions, methods, sources, assumptions, and processes should be clear.

A related concept is that of replicability. It is sometimes said in laboratory science that if an experiment cannot be replicated, it is almost as if it did not happen. The parallel in data science is that

any team possessing similar resources (size, time, financial resources, and data availability), having the same goal (the data product), and following the same protocol would arrive at a data product that yielded, with a high degree of agreement, the same information.

Adhering to these two principles requires a culture of documentation. The following should be accurately recorded for any given data product project:

- Objectives (i.e., the ultimate form the data product is expected to have), along with any downstream modification of them
- Data sources
- All key details of the data preparation process
- Any analytical techniques applied to the data to extract actionable information
- Any steps taken in packaging the result into a data product
- The key information conveyed regarding the data product to data product consumers
- All supporting programs, documents, spreadsheets, scripts, training materials, etc., associated with these steps

This level of documentation is a necessary bureaucratic burden. A challenge for data science teams and management will be to develop standardized procedures and processes for this documentation to make it less burdensome. Indeed, the ideal would be a documentation process that is as close to an automatic and unconscious element of data science workflow as possible. Naturally, this documentation process will require some sort of archiving, as well (discussed earlier among the responsibilities of management).

Such documentation will save data products from questions about the integrity or honesty of their process Moreover, even if a data science project never receives outside scrutiny, the need to adhere to the standards of transparency and replicability compels teams never to retreat from the highest standards of rigor, intellectual competition and self-scrutiny. This practice insures that data science is as rigorous and scientifically powerful as possible.

Documentation can also generate internal benefits for data science teams. Data science is conducted in a fast-paced, ever-changing environment. Unfortunately, the limitations of human memory really have not changed. Teams can refer to their own documentation for guidance when they encounter a recurring challenge: there is nothing to be gained from reinventing the wheel.

Ethics of Data Science

In the past several years, concern about privacy and personal security in the age of the Data Tsunami has been mounting. As data scientists, we tend to think of the enormous amount of information in the tsunami as a fantastic opportunity to learn about the world. As citizens, though, many of us are justifiably concerned about how much personal information is captured and made available to the wider world. A lot of this is information we may not be comfortable sharing. With data science teams, a basic common concern arises: How does one ethically exploit the information in the Data Tsunami, when the source of that information is people who may not know that it is being captured,

probably do not know that it could be applied by data science, and may or may not have consented to its use as a data product?

These questions have no easy answers, and a full examination of the topic is beyond the scope of this document. Data scientists walk a fine line between information and privacy, and ethical and nonethical uses of data.

CONCLUSION

Data has value only if it leads to knowledge. Obtaining knowledge from data is a matter of finding appropriate data, analyzing it effectively, and then communicating the findings to prompt an action. If global health professionals cannot find, analyze, or use the data they need, they will be no better off than before the Data Tsunami existed.

Data science offers an approach to creating actionable information from the Data Tsunami. The discipline's principles of practice make it possible to find useful data efficiently, both in traditional realms and in new ones. Once this data is found, data science's innovative analysis techniques can uncover deeper levels of the data's meaning, and advances in data visualization and communication can move decision makers from recognition into action.

REFERENCES

Abelson, B., Varshney, K., & Sun, J. (2014). Targeting direct cash transfers to the extremely poor. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Retrieved from http://dl.acm.org/citation.cfm?id=2623335

Brown, D. How computer modelers took on the Ebola outbreak." IEEE *Spectrum* blog, May 28, 2015. Retrieved from http://spectrum.ieee.org/computing/software/how-computer-modelers-took-on-the-ebola-outbreak

Data Revolution Group. (2014). *A World That Counts: Mobilising the Data Revolution for Sustainable Development*. Available at: http://www.undatarevolution.org/report/

Internet World Stats. [n.d.] Africa 2015 population and Internet users statistics for 2015 Q3. Website. Retrieved from http://www.internetworldstats.com/stats1.htm

McLeod, D. Feature phone sales fast dying in Africa. TechCentral, July 12, 2015. Retrieved from http://www.techcentral.co.za/feature-phones-fast-dying-in-africa/58129/.

Raftree, L., & Bamberger, M. (2014). Emerging opportunities: monitoring and evaluation in a tech-enabled world. East Sussex, United Kingdom: Itad, Ltd. Retrieved from https://www.rockefellerfoundation.org/report/emerging-opportunities-monitoring/

Sarasohn-Kahn, Jane. Consumer-generated data in a big data world: report from the California Healthcare Foundation. Health Affairs Blog, August 20, 2014. Retrieved from http://healthaffairs.org/blog/2014/08/20/consumer-generated-data-in-a-big-data-world-report-from-the-california-healthcare-foundation/

United Nations Children's Fund. (2013). Tracking anti-vaccination sentiment in Eastern European social media networks. New York, NY: United Nations Children's Fund. Retrieved from http://www.unicef.org/ceecis/Tracking_anti-vaccine_sentiment_in_Eastern_European_social_media_networks.pdf

Worldpop. (n.d.) Ebola. Retrieved from http://www.worldpop.org.uk/ebola/

MEASURE Evaluation
University of North Carolina at Chapel Hill
400 Meadowmont Village Circle, 3rd Floor
Chapel Hill, NC 27517 USA
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. TR-16-143

USAID
FROM THE AMERICAN PEOPLE

MEASURE
Evaluation