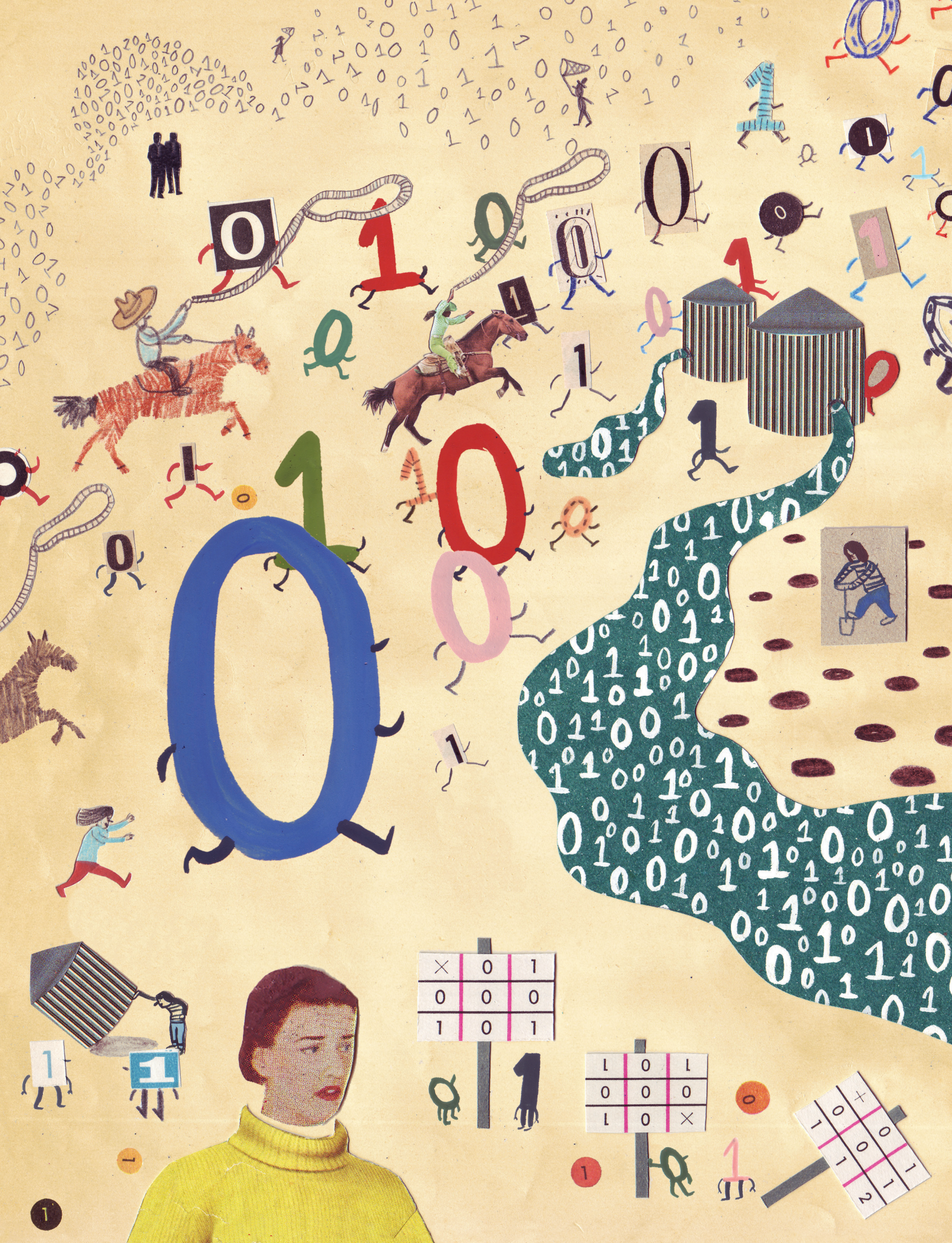
Stanford SOCIAL INNOVATION REVIEW

Big Data for Social Innovation

By Kevin C. Desouza & Kendra L. Smith

Stanford Social Innovation Review
Summer 2014


× 0 1
0 0 0
1 0 1
1 0 1
0 0 0
1 0 ×
+ 0 1
1 0 1
1 1 2

Nonprofits and other social change organizations are lagging their counterparts in the scientific and business communities in collecting and analyzing the vast amounts of data that are being generated by digital technology. Four steps need to be taken to improve the use of big data for social innovation.

Big Data for Social Innovation

By KEVIN C. DESOUZA & KENDRA L. SMITH

Illustration by NICK WHITE

According to IBM, about 2.5 quintillion bytes of data are created every day—enough to fill about 57.5 billion 32 GB iPads daily. Some of these data are gathered by scientific instruments measuring winds, temperatures, and currents around the world. Other data are captured by computers tracking bond sales, stock trades, and bank deposits. And other data are input by police officers, probation officers, and welfare administrators. All of the data, however, are simply that—data—until they are analyzed and used to inform decision-making. What will the weather be like next week? What are the most lucrative investment opportunities? Which neighborhoods should be receiving more social services?

The term "big data" is used to describe the growing proliferation of data *and* our increasing ability to make productive use of it. A myriad of big data projects have been undertaken in scientific domains. For instance, in 2012, pharmaceutical company Merck found through data analysis that allergens would probably lie dormant throughout March and April 2013 because of unseasonably cold weather, followed by a sudden May warm-up that would cause pollen to be released at a higher-than-average rate, thus driving the potential need for Merck's allergy medication Claritin. Merck then modified its marketing strategy to capitalize on the high demand for allergy relief. Through partnerships with Walmart, they created personalized promotions based on zip code data to market Claritin to heavily hit areas, resulting in increased revenue.

The business community has also been a heavy user of big data. Each month Netflix collects billions of hours of user data to analyze the titles, genres, time spent viewing, and video color schemes to gauge customer preferences in order to continually update their recommendation algorithms and programming to give the customer the best possible experience.[1] In 2013, Netflix launched its first original series, *House of Cards*, largely using a mix of customer behavior data and analytics to help shape the story. Netflix invested $100 million into the series without testing a pilot or conducting focus groups, instead banking on the success of an earlier BBC production by the same name about UK politics, along with what it had learned about the preferences of its 44 million customers.[2] *House of Cards* has been a great success, bringing in 2 million new subscribers.

Data-driven intelligence has been used successfully in technical and business endeavors, but a very different situation prevails in the social arena. There, a large chasm exists between the potential of data-driven information and its actual use in helping solve social problems. Some social problems can be readily solved using big data, such as using traffic data to help ease the flow of highway traffic or using weather data to predict the next hurricane. But what if we want to use data to help us solve our most human and critical social problems, such as homelessness, human trafficking, and education? And what if we not only want to solve these problems but do so in a way that the solutions are sustainable for the future?

Social problems are often what are called "wicked" problems. Not only are they messier than their technical counterparts, they are also more dynamic and complex because of the number of stakeholders involved and the numerous feedback loops among inter-related components. Numerous government agencies and nonprofits are involved in tackling these problems, with limited cooperation and data sharing among them. Most of these organizations have inadequate information technology resources, compared to their counterparts in the hard sciences who work on technical problems or in business who have ready access to financial, product, and customer information.

Beyond the infrastructural impediments that social sector users of big data face, data itself can be a problem. Oftentimes, data are missing and incomplete, or stored in silos or in forms that are inaccessible to automated processing. Then there are policy and regulatory challenges that need to be faced, such as building data-sharing agreements, ensuring privacy and confidentiality of data, and creating collaboration protocols among various stakeholders tackling the same type of problem.

Whereas there is no doubt that nonprofits, government, and other organizations will continue to invest in big data technologies and programs, questions still remain about how beneficial those investments will turn out to be. The value proposition of big data is clear for tackling complex technical and business problems, but the jury is still out on how well big data can tackle complex social problems.

WHY DATA IS BIG

Data, or individual pieces of information, have been gathered and used throughout history. What's changed recently is that advances in digital technology have significantly increased our ability to collect, store, and analyze data. Consider the US Census Bureau. In 1880, the United States conducted a national census of 50 million people that collected demographic information including age, gender, number of people in the household, ethnicity, birth date, marital status, occupation, health status, literacy, and place of origin. All of this information was logged by hand, microfilmed, and sent to be stored in state archives, libraries, and universities. It took seven to eight years to properly tabulate census data after the initial collection.

In 1890, the Census Bureau streamlined its data collection methods by adopting machine-readable punch cards that could be tabulated in one year. In the most recent US census, conducted in 2010, the bureau used a range of emerging technologies to survey the populace, including geographic information systems, social media, videos, intelligent character-recognition systems, and sophisticated data-processing software.

Today, big data is used to refer to data sets that extend beyond single data repositories (databases or data warehouses) and are too large and complex to be processed by traditional database management and processing tools. Big data can encompass information such as transactions, social media, enterprise content, sensors, and mobile devices.

There are multiple dimensions to big data, which are encapsulated in the handy set of seven "V"s that follow.

Volume:	considers the amount of data generated and collected.
Velocity:	refers to the speed at which data are analyzed.
Variety:	indicates the diversity of the types of data that are collected.
Viscosity:	measures the resistance to flow of data.
Variability:	measures the unpredictable rate of flow and types.
Veracity:	measures the biases, noise, abnormality, and reliability in datasets.
Volatility:	indicates how long data are valid and should be stored.

Although all seven *V*s are increasing, they are not equal. Consider *volume*. The world's collections of data are doubling every 18 months, presenting the public and private sectors with new opportunities to transform information into insight. As the volume of data increases along with the tendency to store multiple instances of the same data across varied devices, the science of information search and retrieval will have to advance.

The most challenging *V* for organizations is *variety*. Organizations have built information systems to tackle data elements in specific categories. The challenge for many organizations is to find economical ways of integrating heterogeneous datasets while allowing for newer sources of data (in origin and type) to be integrated within existing systems. Ensuring that the data collected are of sufficient *veracity* is also critical. Today, because of the proliferation of social networks and social media, much of the data being collected needs to be thoroughly analyzed before decision-making, as the data can be easily manipulated.

KEVIN C. DESOUZA is the associate dean for research in the College of Public Programs; an associate professor in the School of Public Affairs; and the interim director for the Decision Theater in the Office of Knowledge Enterprise Development at Arizona State University.

KENDRA L. SMITH is a doctoral candidate in the School of Community Resources and Development within the College of Public Programs at Arizona State University.

FAILING TO USE BIG DATA

When considering big data in the context of social problems, we arrive at a humbling conclusion: For the most part there is no big data! When it comes to social problems, data are still highly unstructured and largely limited to numbers, rather than other types of data. Take, for instance, human trafficking, a $32 billion global industry that ensnares an estimated 30 million people annually. Although considerable momentum exists to combat the problem, few initiatives have attempted to use big data.

Increasingly, traffickers make use of mobile phones, social media, online classifieds, and other Internet platforms. Data from these technologies could be collected and used to identify, track, and prosecute traffickers, but a few daunting truths remain: The illicit nature of human trafficking makes it difficult to collect primary data, primary data collected from some organizations may be unreliable, and we lack reliable indicators to measure anti-trafficking program and policy success. Furthermore, most information collected on human trafficking is stored in a manner that meets organizational needs, but not global needs. Because of data privacy and security issues, data held by various organizations are seldom shared in raw form, limiting the creation of global, and big, datasets.

Making matters worse, agencies combatting trafficking often compete with each other for scarce resources, whether grants and gifts or recognition from the press and the community. Because of this competition, data sharing between agencies—and even between agencies and the public—is rare. The Polaris Project, for example, has been working to combat human trafficking using a comprehensive approach combining advocacy, client services, technical training and assistance, global programs, and a national resource hotline. Between 2003 and 2006, Polaris provided hotlines for human trafficking survivors to call. In 2007, the US Department of Health and Human Services selected Polaris as the country's first national human trafficking resource hotline. Over the years, Polaris is believed to have logged more than 75,000 calls; nevertheless, access to the data is limited and little is known about its reliability and its sources.

Think what might be done if the Polaris information was opened to the public and integrated with other data sources, such as economic indicators, transportation routes, education statistics, and victim services. Only when the data are aggregated with other data, analyzed, visualized, and made accessible to a multitude of stakeholders will the collection be truly valuable. Only then will the small data have a chance to grow into big data and help us effectively combat human trafficking.

One hopeful sign is that in 2012 Google Giving awarded Polaris and two other international anti-human trafficking organizations $3 million to fund the aggregation of the data collected from their three hotlines and to scale their hotlines into an international hotline. Together, all three organizations have coalesced under the Global Human Trafficking Hotline Network. This is a positive sign, but it is yet to be seen what the fruits of this collaboration will be.

BARRIERS TO CREATING AND USING BIG DATA

There are four principal reasons for the relative lack of structured big data for social problems: Data are buried in administrative systems, data governance standards are lacking, data are often unreliable, and data can cause unintended consequences.

Data are buried in administrative systems | Most organizations collect data to meet operational needs, and those data are often buried in the organization's administrative systems. To overcome this problem, organizations are trying to find ways to build large datasets that can be more widely used. This obstacle needs to be overcome before we begin thinking of connecting datasets across organizations. Take the US health care industry, for example. Inefficient management of big data costs the industry between $100 billion and $150 billion a year in administrative costs. The biggest problem in the health care industry is the sheer volume of health and insurance plans that providers contract and negotiate with to be paid for their services. Each health or insurance plan supports its own system of underwriting, claims administration, provider network contracting, and broker network management—leaving data stored in multiple formats in multiple places. The McKinsey Global Institute estimated that if the US health care industry were to transform its use of big data for more efficiency and quality, the sector could create more than $300 billion in value every year.

> The issues being tackled in the social sector are often more complex than they are in business or science, making the use of big data more difficult.

Data governance standards are lacking | A second challenge in our ability to use big data for social problems is the lack of adequate data governance standards that define how data are captured, stored, and curated for accountability. As a result, large inconsistencies exist and the data being captured are often not readily suitable for analysis. In many cases data need to be transformed before they can be used, and transformation is costly. Analysts often struggle with integrating different datasets because they lack good metadata (data that describe data) and the quality of data is poor. An example of this hindrance is the US government's 2009 initiative, data.gov, to make its vast amounts of data readily available to the public so that nonprofits, businesses, and other organizations can use the data for innovative purposes. The initiative has been hampered by the difficulty of ensuring that the data are in a usable format. Data quality differs heavily from agency to agency, with some agencies, such as the Environmental Protection Agency, releasing data regularly and in machine readable formats, whereas other agencies publish data in difficult-to-manipulate forms such as PDFs or older file formats.[3] The number of government datasets being made publicly available has exploded, but only a handful of these datasets are ever used. The ones that are being used are, not surprisingly, cases where there is good metadata, ease of accessibility, and manipulability.

Data are often unreliable | The abundance of data provides great opportunities to researchers trying to understand and solve social problems, but unfortunately much of the data is unreliable. Simply having a lot of data does not necessarily mean that the data are representative and reliable. For example, in 2011, the Obama

Administration proposed the Keystone XL pipeline project to carry tar sands oil from Alberta, Canada, to Texas. This proposal raised concerns among landowners, farmers, ranchers, and environmentalists who were living in the vicinity of the proposed pipeline. Despite the concerns, the American Petroleum Institute and its oil lobby allies were able to manipulate social media sentiment to show support for the project. They did so by using Twitter to send an inordinate number of tweets to show support for the project, which did not accurately represent the overall public sentiment. The Rainforest Action Network (RAN) discovered this subterfuge, criticizing the oil companies for using fake Twitter accounts to show support for the pipeline project. RAN pointed out a sudden spike in the number (within three minutes on 15 accounts) of tweets favoring the pipeline. RAN gathered evidence that 14 of 15 accounts were phony and the tweets were generated by an automated process.

Data can cause unintended consequences | Big data users can find themselves facing the unintended consequences of exploiting big data with no regard for data quality, legality, disparate data meanings, and process quality.[4] This was the case when public agencies and a newspaper in New York came under scrutiny for releasing information about gun owners. In the wake of the Connecticut school mass shooting, a group of journalists from *The Journal News* in White Plains, N.Y., used the Freedom of Information Act to obtain information regarding gun owners living in the suburbs of Westchester, Rockland, and Putnam counties. The journalists published an article about the licensed gun owners living in the neighborhood and also published an interactive visual map that provided individual gun owners' names and addresses. The information was published to inform the public about who owns firearms, but that information might also assist criminals who could use it to target vulnerable homeowners who do not own guns or to target homeowners who have guns in order to steal them.[5]

THE PROMISE OF MOBILE PHONES

There is one area where nonprofits have begun to make good use of big data: mobile phones. In 2010 more than 5 billion mobile phones were in use, more than 80 percent of them in developing countries.[6] The percentage of people owning mobile phones in Sub-Saharan Africa increased from 32.1 percent in 2008 to 57.1 percent in 2012, and it is expected to rise to 75.4 percent by 2016.[7] This growth has offered people in developing countries better opportunities to improve their quality of life.

For example, Cell Life, a South African organization, created a mass messaging mobile service called Communicate, which reminds patients to take their medications, links patients to clinics, and offers peer-to-peer support services such as counseling and monitoring.[8] Cell Life also developed Capture, a service that makes it possible for health care workers in the field to collect and save information in digital form using their mobile phones.

The rapid proliferation of mobile and Internet usage allows for the collection of unprecedented amounts of information. Most modern mobile phones contain global positioning system technology, which identifies the geographic location of the phone. In addition to location data, mobile phones contain a treasure trove of information, such as call logs, SMS messages, and social media postings. A mobile phone acts as an individual sensor collecting relevant information from its environment, which when aggregated and analyzed with information from millions of other mobile phones can lead to the discovery of important information, which can then be disseminated back to people on the ground via the same mobile phones.

For example, researchers are studying migration movements following disasters as a way to understand the spread of infectious disease. Harvard University epidemiologist Caroline Buckee and her team use location data from mobile phones to understand the patterns of people moving around in Kenya and help stop malaria and other diseases from spreading.

Kenya's western highlands are equipped with thousands of cell-phone towers that transmit data on individuals' phone call and text messaging activity. Researchers found that people making calls and sending text messages from a specific tower were making 16 times more trips away from the area, with significant activity in the malaria hot spot of Lake Victoria. Information on the patterns of human travel collected from mobile phone usage are being used to develop predictive models to further combat malaria in the region.[9]

> With the proliferation of open data platforms, citizens are creating new ideas and products through what has become known as "citizen science."

STEPS TO INCREASE USE OF BIG DATA

Big data has enormous potential to inform decision-making to help solve the world's toughest social problems. But for this to happen, issues relating to data collection, organization, and analysis must first be resolved. The following four recommendations have the potential to create datasets useful for evidence-based decision-making.

Building global data banks on critical issues | The global community needs to create large data banks on complex issues such as human trafficking, global hunger, and poverty. The data bank would have the capacity to hold multiple different data types along with metadata that describes the datasets. For this to happen, multi-sector alliances that promote data sharing on thematic issues need to be created. At the 2012 G-8 Summit, leaders of the world's largest economies and four African heads of state met to discuss and commit to a new phase of efforts to fight hunger and food insecurity. Out of that discussion grew the New Alliance for Food and Nutrition Security, which set its sights on helping 50 million people out of poverty over the next 10 years through sustained agricultural growth. As part of this plan, the New Alliance launched a number of technology- and data-based initiatives. One was the Scaling and Seeds and Other Technologies Partnership, developed to promote commercialization, distribution, and adoption of technologies that would improve seed varieties. The United States' contribution to the New Alliance has been chronicled through the Feed the Future Initiative and website, and it has stayed true to the Alliance's stance on data sharing by establishing Agrilinks.org, a data-sharing platform that is updated consistently. Farmers can tap into Agrilinks.org to read about new agricultural practices or live tweet from

their mobile phones to ask questions of an agriculture expert. USAID is offering open data from the Feed the Future initiative on baseline data pulled from the Bangladesh Integrated Household Survey dataset,[10] baseline surveys of nearly 5,000 households in Ghana that captured indicators outlined by the Feed the Future Initiative[11] and the Women's Empowerment in Agriculture Index.[12]

Engaging citizens and citizen science | Big data is not the sole province of professionals. Citizens can also be enlisted to help create and analyze these datasets. With the proliferation of data through open data platforms, more and more citizens are creating new ideas and products through what has become known as "citizen science." In 2010, the City of London made government data available to the public by opening the London Datastore. Managed by the Greater London Authority, the London Datastore offered citizens the opportunity to view and use raw data released from city agencies and civil servants. Information distributed included data on crime and economics, and real-time data from transit systems. Matthew Somerville, a Web developer, created an online map app of the City of London tube that had more than 250,000 hits in a matter of days. Likewise, Ben Barker, an electronics engineer and cyclist, created a bike map with information pulled from the London Datastore.[13]

Build a cadre of data curators and analysts | Today, not only do we have a shortage of data curators and analysts who can tackle social problems, we have limited avenues for our existing personnel to receive the necessary training and build competencies. For the most part, we have left data science to the sciences and business. The social sciences have often equipped students simply with the basics of statistics. This approach is unacceptable if we are to take advantage of big data. We need to equip students and analysts with the necessary skills to curate data so as to create large datasets. These skills are often found in programs in informatics and the traditional degrees of information and library science. In these programs, students learn about data organization, preservation, visualization, search and retrieval, and use. These are valuable skills that go beyond simply searching the Web for information. In addition to these skills, increasing the capacity for an analyst to think about what is possible with data is critical. Thinking about networked relationships among datasets, and how to uncover latent patterns in datasets, are competencies that need to be developed.

Promoting virtual experimentation platforms | To increase our understanding of how to use big data for tackling social problems, we need to promote more experimentation. Virtual experimentation platforms, which allow individuals to share ideas, interact with others' ideas, and work collaboratively to find solutions to problems or take advantage of opportunities, can bring interested parties together to create large datasets, develop innovative algorithms to analyze and visualize the data, and develop new knowledge. One example is Kaggle, a website where competitions on data analysis are run. Unfortunately, organizations that are tackling social issues seldom participate on these platforms. Virtual experimentation platforms are essential if we are going to move the needle on using big data to tackle social challenges. Initially, these platforms should stimulate competitions to create large datasets on various issues. Competitions that generate large datasets will be critical to help the community realize the challenges associated with the way the social sector is currently operating. Once a couple of datasets are created, we can launch competitions that focus on the predictive analytics and the discovery of novel patterns. The use of open forums such as wikis and discussion groups can help the community share lessons learned, collaborate, and advance new solutions.

THE FUTURE OF BIG DATA

Business and science have shown that big data's merits are undeniable. Social sector organizations must now figure out how they too can incorporate this type of decision-making capability into their operations. The potential for growth and innovation exists, but there are serious obstacles to overcome. The issues that are being tackled in the social sector are in many ways more complex than they are in business or science, making the use of big data that much more difficult. In addition, greater attention must be paid to the rights, privacy, and dignity of their constituents.

In spite of these obstacles, progress is being made. Public sector agencies have made it clear that data are an important element of social innovation. Institutions such as the US government and the World Bank have made their data available to the public for mining and further use. Individuals are using the data to create innovations, mainly apps, to address a particular social problem.

Organizations have been created to help make better use of big data for social problems. Data Without Borders, for example, matches scientists and statisticians with nonprofits for pro bono data work to help overcome the shortage of technology personnel capable of handling big data projects. Globally, the world's actors are making efforts to use open data and big data to develop solutions to social problems in innovative and collaborative ways. Progress is being made, but the chasm must still be crossed. It is a challenge worth overcoming. ■

NOTES

1 Phil Simon, "Big Data Lessons from Netflix," Wired.com, Mar. 11, 2014. http://www.wired.com/insights/2014/03/big-data-lessons-netflix/

2 Mark Sweney, "Netflix Gathers Detailed Viewer Data to guide its Search for the Next Hit," TheGuardian.com, Feb. 23, 2014. http://www.theguardian.com/media/2014/feb/23/netflix-viewer-data-house-of-cards

3 Joseph Marks, "Data Dreams," *Government Executive*, Feb. 1, 2012. http://www.govexec.com/magazine/nextgov/2012/02/data-dreams/40992/

4 Marcus R. Wigan & Roger Clarke, "Big Data's Big Unintended Consequences," *Computer*, vol. 46, no. 6, June 2013, pp. 46-53.

5 Unkown, "White Plains Police Investigate Possible Link between Burglary, Journal News Gun Permit Map," NewYork.CBSLocal.com, Jan. 14, 2013. http://newyork.cbslocal.com/2013/01/14/white-plains-police-investigate-possible-link-between-burglary-journal-news-gun-permit-map/

6 Emmanuel Letouzé, "Big Data for Development: Challenges & Opportunities UN Global Pulse," UNGlobalPulse.org, May 2012. http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

7 Madanmohan Rao, "Mobile Africa Report 2012: Sustainable Innovation Ecosystems," MobileMonday.net. http://www.mobilemonday.net/reports/MobileAfrica_2012.pdf

8 http://www.cell-life.org/systems/communicate/

9 David Talbot, "Big Data from Cheap Phones," MITTechnology Review.com, April 23, 2013. http://m.technologyreview.com/featuredstory/513721/big-data-from-cheap-phones/

10 The Feed the Future Bangladesh Integrated Household Survey dataset, Data from 2011-2012, USAID.gov, April 23,2013. http://www.usaid.gov/developer/FTFBangladesh

11 Ghana Feed the Future Baseline Survey Dataset, Data from 2011, USAID.gov, April 23, 2013. http://www.usaid.gov/developer/FTFGhana

12 Feed the Future: Women's Empowerment Agricultural Index, Data from 2011-2012, USAID.gov, April 23, 2013. http://www.usaid.gov/developer/WEAI

13 Mark Halper, "London: Turning Access into Apps," Time.com, Jan. 6, 2011.