GOVLAB

THE POTENTIAL OF SOCIAL MEDIA INTELLIGENCE TO IMPROVE PEOPLE'S LIVES

Social Media Data for Good

September 24, 2017

By Stefaan G. Verhulst and Andrew Young

THE POTENTIAL OF SOCIAL MEDIA INTELLIGENCE TO IMPROVE PEOPLE'S LIVES

Social Media Data for Good

September 24, 2017

By Stefaan G. Verhulst and Andrew Young*

COVER IMAGE: YOLANDA SUN

* Stefaan G. Verhulst is Co-Founder and Chief Research and Development Officer of The Governance Lab
Andrew Young is Knowledge Director at The Governance Lab

TABLE OF CONTENTS

Acknowledgements

With this report we not only sought to write about data collaboratives around social media data but wanted to also actually collaborate with social media companies in writing the report. We were therefore delighted to have Facebook support the report both financially and intellectually. In particular we benefitted greatly from engaging with Molly Jackman, Chaya Nayak, Nicole Fulk and Tarunima Prabhakar at Facebook. Juliet McMurren, Audrie Pirkl and Hendrick Townley contributed tremendously to the case studies and broader analysis included in the report. The report also benefited from Akash Kapur's invaluable editorial support. Finally, thank you to the group of expert peer reviewers who provided important input on a pre-published draft: Natalia Adler, Jos Berens, Joseph Jerome, Patricia Muylle, Matt Stempeck, Stuti Saxena, Josje Spierings, and Shenja Van Der Graaf.

USING DATA TO BETTER UNDERSTAND THE WORLD AROUND US

People turn to Facebook to share with friends and family, find new businesses, and even connect with their elected representatives. We want to build products that help people feel supported, safe, informed, engaged and included - creating a community that works for everyone. One of the best ways to build better features and experiences is to use data science to quickly and accurately understand what is happening on Facebook.

At Facebook, most research is focused on making sure we build products that are responsive to the needs of our community — not just based on our instincts or intuitions. Data is at the center of the products that provide value to people on Facebook every day. Facebook's data, reach, and research tools can also help businesses, organizations, and governments make better decisions that improve lives.

That's why we're excited about this report – it not only captures the successes that data and technology have had to make the world a better place, but also it highlights some of the challenges the data and technology community must address to generate an even larger impact.

At Facebook, we partner with trusted organizations globally to create data-driven solutions to help solve the world's most pressing problems - on and offline. In addition to partnering with the right organizations at the right time, our approach utilizes statistics and data processing to preserve privacy. These methodologies filter data signal from noise and make insights more effective. Finally, we have invested in internal privacy and review systems to build an infrastructure for ensuring ethical and privacy preserving data sharing processes.

These are just the first steps that we are taking. We'll continue to learn how to best contribute to efforts to build social value through data.

Chaya Nayak
Facebook Public Policy Research Manager

THE POTENTIAL OF SOCIAL MEDIA INTELLIGENCE TO IMPROVE PEOPLE'S LIVES

By Stefaan G. Verhulst and Andrew Young

Executive Summary

The twenty-first century will be challenging on many fronts. From historically catastrophic natural disasters resulting from climate change to inequality to refugee and terrorism crises, it is clear that we need not only new solutions, but new insights and methods of arriving at solutions. Data, and the intelligence gained from it through advances in data science, is increasingly being seen as part of the answer. This report explores the premise that data—and in particular the vast stores of data and the unique analytical expertise held by social media companies—may indeed provide for ***a new type of intelligence that could help develop solutions to today's challenges.***

In this report, developed with support from Facebook, we focus on an approach to extract public value from social media data that we believe holds the greatest potential: data collaboratives. Data collaboratives are an emerging form of public-private partnership in which actors from different sectors exchange information to create new public value. Such collaborative arrangements, for example between social media companies and humanitarian organizations or civil society actors, can be seen as possible templates for leveraging privately held data towards the attainment of public goals.

THE PROMISE OF DATA COLLABORATIVES

Existing research on data collaboratives is sparse, but a number of recent examples show how social media data can be leveraged for public good. These include Facebook's sharing of population maps with humanitarian organizations following natural disasters; predicting adverse drug reactions through social media data analysis in Spain; and the city of Boston's use of crowdsourced data from Waze to improve transportation planning. These examples and 9 additional cases are discussed in the full report.

By assessing these examples, we identify **five key value propositions** behind the use of social media data for public goals:

1. SITUATIONAL AWARENESS AND RESPONSE

Data held by social media companies can help NGOs, humanitarian organizations and others better understand demographic trends, public sentiment, and the geographic distribution of various phenomena. In doing so, data contributes to improved situational awareness and response.

Case Studies:

- Facebook Disaster Maps
- Tracking Anti-Vaccination Sentiment in Eastern European Social Media Networks
- Facebook Population Density Maps

2. KNOWLEDGE CREATION AND TRANSFER

Widely dispersed datasets can be combined and analyzed to create new knowledge, in the process ensuring that those responsible for solving problems have the most useful information at hand.

Case Studies:

- Yelp Dataset Challenge
- MIT Laboratory for Social Machines' Electome Project
- LinkedIn's Economic Graph Research Program

3. PUBLIC SERVICE DESIGN AND DELIVERY

Data Collaboratives can increase access to previously inaccessible datasets, thereby enabling more accurate modelling of public service design and helping to guide service delivery in a targeted, evidence-based manner.

Case Studies:

- Facebook Future of Business Survey
- Easing Urban Congestion Using Waze Traffic Data
- Facebook Insights for Impact Zika

4. PREDICTION AND FORECASTING

New predictive capabilities enabled by access to social media datasets can help institutions be more proactive, putting in place mechanisms based on sound evidence that mitigate problems or avert crises before they occur.

Case Studies:

- Tracking the Flu Using Tweets
- Predicting Floods with Social Media Metatags
- Predicting Adverse Drug Events by Mining Health Social Media Streams and Forums

5. IMPACT ASSESSMENT AND EVALUATION

Access to social media datasets can help institutions monitor and evaluate the real-world impacts of policies. This helps design better products or services, and enables a process of iteration and constant improvement.

Case Studies:

- Sport England's #ThisGirlCan
- Using Twitter Data to Analyze Public Sentiment on Fuel Subsidy Policy Reform in El Salvador
- Using Twitter to Measure Global Engagement on Climate Change

RISKS

RESPONSIBLY SHARING SOCIAL MEDIA DATA

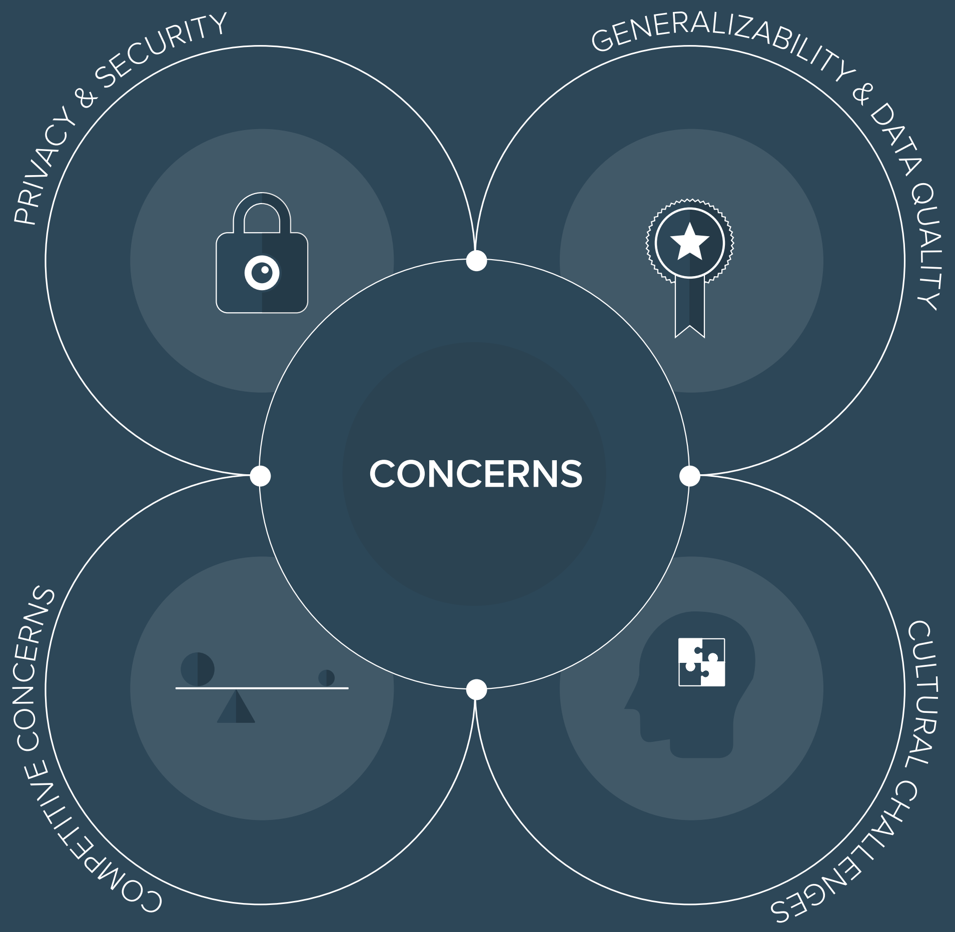


FIGURE 1

Despite the potential of data collaboratives, companies and public organizations often have concerns about sharing data. Many of these concerns are legitimate: data sharing is not without risks and challenges. We identify four key risks and challenges, and discuss ways to mitigate them. These include:

PRIVACY AND SECURITY

The most common concern expressed by individuals and companies involves the concern that sharing information may result in disclosing personally or demographically identifiable information, which may create privacy and/or security violations. Such concerns are not only natural, but very important: data sharing must not result in any dilution of protections for individuals, many of whom might not even be aware that the data was collected about them in the first place.

COMPETITIVE CONCERNS

Companies are often concerned that sharing data—usually without charge—will threaten their commercial interests or affect their competitive advantage. While such concerns are important to address, our research into the field of cross-sector data-sharing[1] suggests that this view is based on a false, zero-sum understanding of data collaboration and its potential. There are often methods of balancing competitive risk with data sharing for public good – such as aggregating data or sharing insights from datasets rather than the raw data.

GENERALIZABILITY, DATA BIAS, AND QUALITY

A key concern when using social media data involves the level of representativeness or data bias. Social media data is often gathered from a particular demographic subset, possibly ignoring so-called "data invisibles"—individuals, often from vulnerable communities, who are unrepresented in private or public datasets. For these reasons, caution needs to be exercised in extrapolating general observations from such data.

BARRIERS TO A CULTURE OF DATA SHARING AND COLLABORATION

Our exploration indicates that one of the chief obstacles to more widespread data sharing—and one that may underlie other concerns—is a lack of understanding about the benefits of sharing social media data, and a lack of comfort and familiarity with such strategies. Social media companies today operate in a rapidly changing environment where notions of collaboration, sharing, and mutual benefit are far more widely accepted as philanthropic and commercial propositions. Embedding these values in business operations is one of the central challenges—and greatest opportunities—facing social media companies today.

TOWARD DATA RESPONSIBILITY

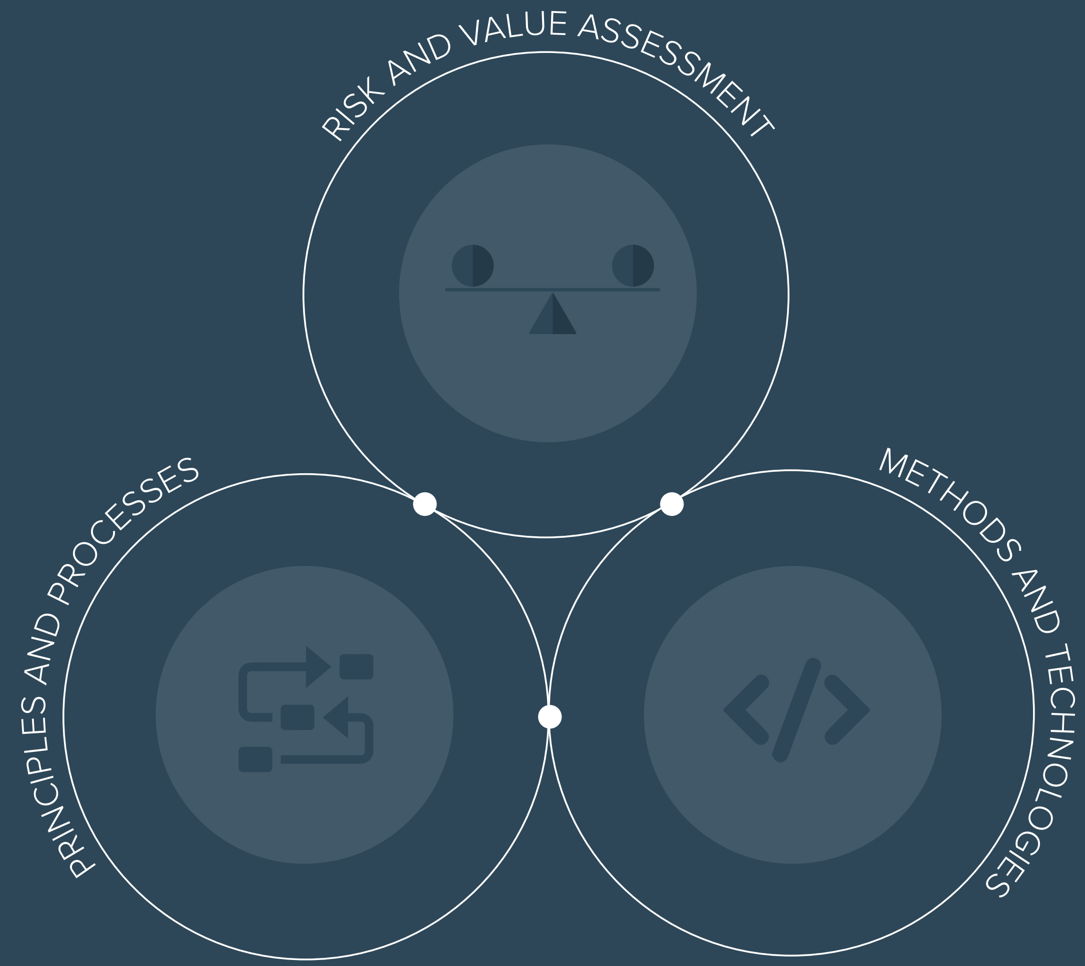


FIGURE 2

Some social media corporations and organizations, including Facebook, the International Committee of the Red Cross, and United Nations Global Pulse, have started to explore frameworks to share data responsibly. For such efforts to succeed, they will likely need to include the following four elements to assure concerns are addressed meaningfully and legitimately:

RISK AND VALUE ASSESSMENTS

Risks—including privacy, ethical and commercial concerns—exist across the social media data lifecycle, and include inaccurate, non-representative data entry during collection; insufficient, outdated, or inflexible security provisions during processing; incompatible cultural or institutional norms or expectations during sharing; aggregation or correlation of incomparable datasets during analysis; and controversial or incongruous data usage. Assessing such risks, and the risks of not sharing data with the potential for transformative impact, is an important task for all data collaborative efforts.

NEW METHODS AND TECHNOLOGIES

Data responsibility frameworks will also need to evolve beyond written policies to create tools and technologies that enable flexibility and context-specific implementation. Data responsibility decision trees, such as the Center for Democracy & Technology's Digital Decision Tool, can be used to translate principles into a series of questions. Consideration of and experimentation with differential privacy could also help to limit privacy risks. Similarly, a transparency report showing with whom data is being shared and toward what public benefit could help allay concerns about government misuse of private-sector data assets.

PRINCIPLES AND PROCESSES

Having a set of principles that can determine if and how social media data can be shared is important for accountability and effectiveness. While frameworks exist, data responsibility principles will need to be updated or newly designed to address emergent risks. Initiatives like the Signal Code and the Handbook on Data Protection in Humanitarian Action are important, but are limited to specific contexts and sectors. More broadly applicable and mutually agreed upon principles could lead to greater uptake and impact.

Principles and assessment frameworks have no value if there is no process in place to apply them or oversee their implementation. Including governance processes is a key element to achieve data responsibility. Such processes should be transparent and participatory, while being flexible and responsive to different needs and contexts. For example, to accommodate collaborative research using social media data, Facebook designed an innovative review process that involved in-house training, different stages of review, and the application of evaluation criteria to determine whether to go ahead.

REALIZING THE POTENTIAL OF SOCIAL MEDIA INTELLIGENCE THROUGH DATA COLLABORATION

From our discussion, we derive a series of recommendations for developing data collaboratives, which we have grouped into four broad categories. Taken together, these form a roadmap for companies or organizations considering using data for public ends.

STEWARDS

Social media corporations should consider themselves, and act as, the standard bearers for a new corporate paradigm of *data stewardship*. This coinage represents a move away from the concept of data as something to be owned and towards stewardship of data as a public good.

Social media companies should pioneer the role or position of Data Stewards within their organizations. Such individuals (or offices) would be tasked with matching and coordinating the demand and supply of social media data in a trusted way. These new positions could provide models for other companies considering the use of private data for public goals.

Among their other roles, **Data Stewards could help develop new coordinating mechanisms to unlock corporations' supply of social media data sets with potential public interest value.** Such mechanisms must include a due process to respond to data requests; a system for filtering or prioritizing certain kinds of information; and a method to ensure that the data being released matches public needs and demands.

EVIDENCE

A more detailed repository of case studies should be established to document impact and practice. Such a repository, which could build on the foundation offered by the case studies presented here, would highlight best practice in value propositions, technical arrangements, and legal frameworks for data collaboratives and give strategies for measuring impact.

Existing metrics and systems of measurement may not be sufficient to capture the true impact and value of data collaboratives. **Companies and organizations involved in data sharing can together help develop better metrics of value and impact**, and indeed, entirely new definitions of success.

We also **need a more granular understanding of how available data matches information needs**. Such an understanding can be achieved through data audit methodologies and tools.

METHODS

In order to scale the use of data collaboratives and realize the potential of the underlying data, **we need a better understanding and systematization of the methods for their deployment**. Organizations and researchers should mine existing data collaborative experiments, such as those to be included in the repository described in the previous section, for examples of successful practice.

Lessons and observations from this can be translated and shared as a toolkit or roadmap for corporations considering sharing data. They would form the basis of a published (and ideally, openly available) "Best Practices" kit that would constantly be updated and improved, based on the latest evidence.

We also need a better understanding of new techniques used by Data Collaboratives to analyze and seek insight from data. The field is built on a number of emergent techniques—including natural language processing, neural networks, computational social science, network science, sentiment analysis, data-mining and machine-learning—each of which needs further study in the context of data sharing .

Many nonprofit and civil society organizations lack the necessary expertise to apply these techniques within Data Collaboratives. **Corporate data sharing initiatives need also to consider sharing their expertise in handling data**—for example, through training initiatives, educational programs, informal peer- and practitioner-mentoring mentoring setups, and the development of affordable, user-friendly tools.

MOVEMENT

Data sharing cannot be thought of in isolation from other activities or environments. Much **more thought and energy needs to be directed toward developing a network for collaboration, supporting ecologies and a platform for sharing**. What's required is a Data Collaboratives movement.

To that end, **efforts should be made to bring together various actors from the social media data community**—including those currently acting as data stewards—at dedicated convenings to share lessons learned; identify pain points; and develop common solutions, procedures, and practices.

Corporations could help facilitating such convenings by providing a venue (virtual or physical) where data providers and users can build knowledge in an emerging field and co-create ideas and insights.

Creating a supporting ecology means interacting with and nurturing all actors in that ecology. As mentioned above, the practice of **data sharing needs to move beyond data per se to include a wider set of skills, expertise and knowledge areas**.

As the data collaboratives and social media intelligence movement continues to take shape, **engagement with the populations whose attributes and behaviors often make up the data held by social media companies will be key**.

In addition to engaging with other skills, there also **needs to be broader engagement with other actors whose decisions can play an important—perhaps even determining—role in the success of data collaboratives**. Such actors could include policymakers, regulators, and potential funders, both for-profit and nonprofit (philanthropic).

> …we look to the potential of leveraging data to provide a new type of intelligence that could help develop global solutions to global challenges.

THE POTENTIAL OF SOCIAL MEDIA INTELLIGENCE TO IMPROVE PEOPLE'S LIVES

By Stefaan G. Verhulst and Andrew Young

I. INTRODUCTION

Today's Problems Require New Intelligence

Every century throws up its particular challenges, but the twenty-first century is already shaping up to be one of the most demanding in recent history. From the growing severity of climate change, to increases in inequality and concerns over automation and joblessness, to burgeoning refugee and terrorism crises: the world faces a complex and often inter-dependent number of problems that have no easy answers.

We require not only new solutions, but ***new methods*** of arriving at solutions. It is clear that we need new forms of intelligence that will radically change how organizations and individuals innovate and search for insight. Data and the increasing "datafication" of everyday life represent an important piece of the puzzle. We live in an era where unprecedented amounts of information are being generated. It is widely recognized now that within these vast stores of information, dispersed across geographies and organizations, there exists the potential for truly transformative forms of collaboration and intelligence.

One of the central arguments of this report is that collaboration—across sectors, across borders, and across organizations—will offer the best hope of a pathway through our current dilemmas. The complexity of our problems means that no single actor is likely to be able to solve them. With globalization and greater interconnectivity, our problems are shared and collective, and it is only natural that our solutions must be collective, too. At its broadest level, this report is concerned with the search for precisely such a collaborative approach; in particular, we look to the potential of leveraging data to provide **a new type of intelligence that could help develop global solutions to global challenges.**

THE UBIQUITY AND POTENTIAL OF SOCIAL MEDIA DATA

In considering the potential of data, it is natural to turn to the role of the large social media companies—Facebook, Twitter, Google, Microsoft, Pinterest, among many others—that increasingly inform our social, economic, cultural and political lives. These social media platforms have over the past decade or so changed the nature of human interactions, and perhaps even what it means to be human.[2] The constant streams of data they generate, often in real time, coupled with advances in data analysis and processing, offer an unprecedented insight into human behavior and contemporary life. Consider some of the following figures:

- There are about 5.05 billion mobile users[3], and 2.78 billion active mobile social users.
- Facebook has 2.046 billion monthly active users[4], and over 50 percent of its users[5] engage on Facebook at least once a day.
- Facebook's "like" button is pressed 2.7 billion times[6] each day.
- Twitter processes about 6,000 tweets[7] every second, or half a billion (500,000,000) public tweets per day.[8]
- Above all, there are an estimated 3.02 billion social media users, or about 40 percent of the world's population,[9] sharing and creating a constant stream of information

> New forms of inter-sectoral collaboration across the data value chain—what we call a data collaborative—are required in order to leverage social media data for the public good and provide for actionable social media intelligence that can improve people's lives.

As much as the sheer volume, it is the diversity of this data that is striking, and potentially transformational. Large social media platforms have created arguably the largest, richest and most dynamic evidence base[10] for human life ever recorded in the course of history. Yet the collection of the data itself won't provide value and insight. Several steps determine the social media data value chain (see picture). Social media technologies provide an unprecedented amount of data at a time when, more importantly, society's ability to process, analyse and leverage massive amounts of data are also being transformed.

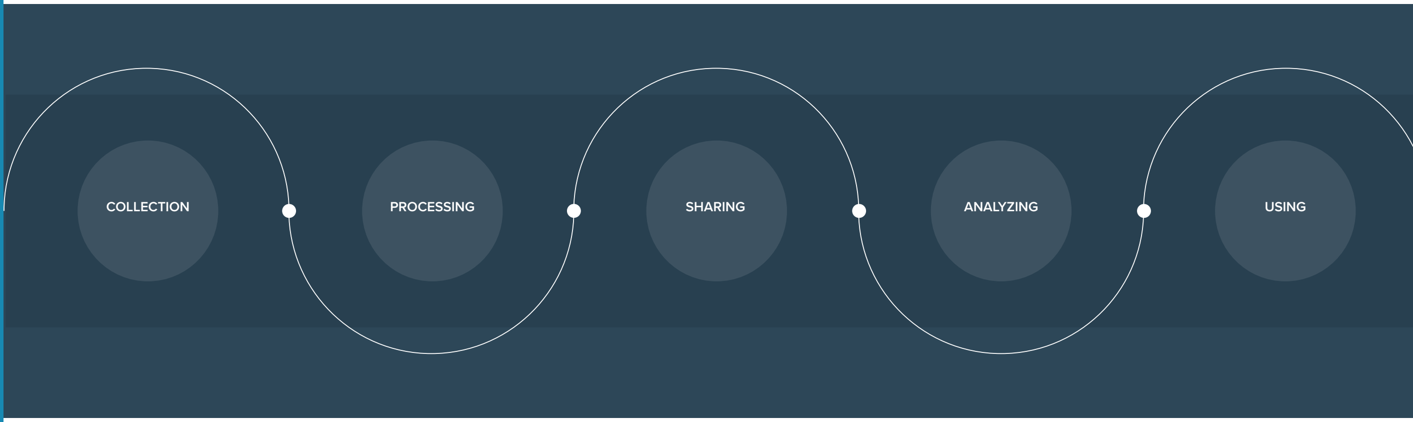


FIGURE 3

While justified debate and some controversy exist surrounding the way companies collect, analyze and use personal data (for instance, to microtarget advertisements), our focus in this report is on the design of a social media data value chain that can create public value. We argue that social media data—if properly processed, analyzed and used, with attention to concerns like privacy and security—could contain within itself answers to some of our most seemingly intractable questions, and provide for this new type of intelligence necessary to address today's problems.

Yet in order to unlock this potential, new avenues need to be created so that the widest range of expertise, intelligence and analytical capacity is brought to bear on the underlying information. Much social media data today is held in virtual silos, as the private property of companies that have generated it. This data will only meet its transformational potential if those silos are unlocked, and if the data is leveraged together with a wider community of researchers, policy makers, international organizations, NGOs and others. New forms of inter-sectoral collaboration across the data value chain—what we call a data collaborative—are required in order to leverage ***social media data*** for the ***public*** good and provide for actionable social media intelligence that can improve people's lives.

THE PROMISE—AND CHALLENGES—OF SOCIAL MEDIA INTELLIGENCE THROUGH DATA COLLABORATIVES

In this report, developed in collaboration with Facebook, we focus on one particular form of collaboration that we believe holds great promise: data collaboratives.[11] Data collaboratives are an emerging and increasingly common form of public-private partnership[12] in which actors from different sectors exchange information to create new public value. Such collaborative arrangements, for example between social media companies and humanitarian organizations or civil society actors, can be seen as possible templates to bring about the social media intelligence we describe above. Data collaboratives are often built on more than simply the exchange of data; with the cross-sector exchange of expertise, knowledge, and resources also playing a key role in achieving success.

With the exception of certain data science conferences and journals, little analysis exists on emerging practices of leveraging social media data for creating public value. Most of the attention on how to use private data for public good is focused on mobile phone[13] or satellite data and the resulting insights, or the use of data manually scraped from social media APIs with no direct collabroation. This report seeks to rectify this and provide a broad mapping across social media platforms and regions of how social media data is being leveraged to generate new insights and intelligence that can improve people's lives. It is not meant to be fully comprehensive or academic, yet seeks to provide a frame of analysis and a conversation on the potential and challenges of social media intelligence.

Our discussion below (*Sections II and III*) highlights a number of specific examples of efforts at leveraging social media data, through data collaboratives and other means, from around the world. These examples have been curated to illustrate the diversity of value propositions, data science techniques and sharing methods, and to be geographically, corporation and sector inclusive. The selection does not represent the full spectrum of private data sharing but seeks to provide a basis to consider the potential of social media intelligence to improve people's lives.

While this report considers the positive potential of data collaboratives, we recognize that data sharing is not always easy, and comes with a number of risks and often legitimate concerns. These include doubts over the representativeness of data, and worries over consumer privacy and information security. Companies that have collected and possess data[14] may also have competitive concerns about data sharing. We examine some of these concerns further below in *Section III* where we also discuss why sharing data is not only in the public interest,[15] but often in the interests of companies that own the data too. We conclude with some steps that can be taken to incentivize private data holders both at an organizational and individual level to develop pro-active data sharing cultures.

II. SOCIAL MEDIA DATA COLLABORATIVES

Creating Social Media Intelligence to Improve People's Lives

In January 2015, heavy rains triggered unprecedented floods in Malawi. Over the next five weeks, the floods displaced more than 230,000 people[16] and damaged over 64,000 hectares of land.[17] Almost half the country was labelled a "disaster zone"[18] by Malawi's government.

As the humanitarian crisis unfolded, relief agencies such as the Red Cross[19] were faced with the daunting task of allocating aid and resources to places that were virtually unrecorded by the country's mapping data, and thus rendered almost invisible. Humanitarian workers struggled to navigate in many of the most affected areas,[20] and one result was that aid did not necessarily reach those most in need.

To help prevent similar knowledge gaps in the future, researchers, volunteers and humanitarian workers in Malawi and elsewhere, are turning to Facebook. In 2016, as part of its "Missing Maps" project, the Red Cross accessed Facebook's rich population density data to find and map people who were critically vulnerable to natural disasters and health emergencies, but remained unrecorded in existing maps.

During local Mapping Parties,[21] volunteers in Malawi used Facebook's satellite and population data, in addition to other satellite imagery, to trace roads, houses, and water points across Malawi's communities. Two years later, Missing Maps in collaboration with Facebook has identified more than 2 million people in Malawi,[22] allowing aid and relief organizations to better plan projects in Malawi's disaster prone areas.

Disasters kill nearly 100,000 and affect or displace 200 million people[23] annually. As climate change is expected to increase the frequency and severity of disasters in the near future, leveraging social media data, crowdsourcing and other means will only become more important.

The Malawi partnership is just one manifestation of the concept of social media data collaboratives—a new form of collaboration beyond the public-private partnership model, in which participants from different sectors— including social media companies, research institutions, and government agencies—can exchange data to help solve public problems.

Such collaboratives are emerging in a number of sectors and areas, including health, disaster response, education, poverty alleviation and more. The Malawi case is an example of a collaborative that sought to establish enhanced situational awareness and subsequently improved response, one of a number of value propositions underlying social media data collaborative experimentation.

In what follows we will explore several similar examples where social media data was used to generate new kinds of intelligence for the pursuit of improving people's lives. We consider these examples with reference to the following **five value propositions:**

SITUATIONAL AWARENESS AND RESPONSE

In addition to the Malawi Missing Maps discussed above, we focus on three initiatives aimed at improving situational analysis:

- Facebook Disaster Maps
- Tracking Anti-Vaccination Sentiment in Eastern European Social Media Networks
- Facebook Population Density Maps

KNOWLEDGE CREATION AND TRANSFER

By bringing more and more diverse datasets to bear, the following efforts seek to fill knowledge gaps and ensure that those responsible for solving problems have the most useful information at hand:

- Yelp Dataset Challenge
- MIT Laboratory for Social Machines' Electome Project
- LinkedIn's Economic Graph Research Program

PUBLIC SERVICE DESIGN AND DELIVERY

Access to previously inaccessible datasets can enable more accurate modelling of public service design and guide service delivery in a targeted, evidence-based manner. Examples include:

- Facebook Future of Business Survey
- Easing Urban Congestion Using Waze Traffic Data
- Facebook Insights for Impact Zika

PREDICTION AND FORECASTING

New predictive capabilities enabled by access to social media datasets can help institutions be more proactive, putting in place mechanisms based on sound evidence that mitigate problems or avert crises before they occur. We examine the following examples:

- Tracking the Flu Using Tweets
- Predicting Floods with Flickr Metatags
- Predicting Adverse Drug Events by Mining Health Social Media Streams and Forums

IMPACT ASSESSMENT AND EVALUATION

Access to social media datasets can help institutions monitor and evaluate the real-world impacts of policies and facilitate product- or service-enhancing iteration. For example:

- Sport England's This Girl Can
- Using Twitter Data to Analyze Public Sentiment on Fuel Subsidy Policy Reform in El Salvador
- Using Twitter to Measure Global Engagement on Climate Change

TAXONOMY OF DATA COLLABORATION

For all the growing attention corporate data sharing has recently been receiving, it remains very much a fledgling activity. Much remains to be defined and understood, and there has been little rigorous analysis of different ways of sharing. A recent survey of the landscape[24] **conducted by the GovLab did, however,** identify six main categories of activity along a **spectrum of openness and collaboration**:

Data Cooperatives or Pooling—corporations and other important data-holders group together to create "data pools" with shared data resources;

Prizes and Challenges—corporations make data available to qualified applicants who compete to develop new apps or discover innovative uses for the data;

Research Partnerships—corporate data is often shared with universities and other academic organizations giving researchers access to consumer datasets and other sources of data to **analyze** social trends;

Intelligence Products—shared (often aggregated) data is used to build a tool, dashboard, report, app or another technical device to support a public or humanitarian objective;

Trusted Intermediaries—corporations share data with a limited number of known partners. Companies generally share data with these entities for data analysis and modelling, as well as other value chain activities; and

Application Programming Interfaces (APIs)—APIs allow developers and others to access data for testing, product development, and data analytics. Here we group instances where social media data was manually scraped by data users together with more open APIs since both data access mechanisms achieve many of the same benefits. The more manual approach, however, does lack some important advantages of more direct collaboration (e.g., cross-sector knowledge sharing) and more openly accessible data (e.g., minimal time and resources required to access data).

DATA COOPERATIVES ACCESSIBILITY MATRIX

This diagram shows how accessible different types of data collaboratives are.
For example, data cooperatives or pooling are less accessible (<) but have more (>) collaboration.

	ACCESSIBILITY	COLLABORATION
DATA COOPERATIVES OR POOLING	<	>
PRIZES AND CHALLENGES	>	=
RESEARCH PARTNERSHIPS	<	>
INTELLIGENCE PRODUCTS	<	<
TRUSTED INTERMEDIARIES	<	>
APPLICATION PROGRAMMING INTERFACES (APIS)	>	<

FIGURE 4

JAMES SUTTON

Both the general categories data collaborative types and the specific character of individual data collaboratives exist on dual spectrums of data accessibility and direct collaboration. If a corporation provides an API (or a researcher scrapes data from a particular platform), for instance, accessibility or openness of the data is high but direct collaboration between the data provider and the data user is limited. A trusted intermediary arrangement, on the other hand, might see data remain inaccessible for the vast majority of society, but high levels of collaboration put that data into the hands of specific public entities in a position to act on it.

Our sample of case studies discussed in this report, while not perfectly representative or comprehensive, provides some early indication of which types of data collaboratives tend to be structured to leverage social media data. The current field of practice seems to be focusing primarily on the creation of intelligence products and the leveraging of trusted intermediaries to create new public value with social media data. We also see a wide array of researchers using social media data drawn from APIs or manually scraped from platforms to create new insights. Data pools, on the other hand, are less prevalent—which is not surprising given concerns related to competitive advantage and privacy.

Similarly, our sample shows that social media platforms are enabling data collaboratives in a few sectors in particular. Crisis response, public health and urban planning, in particular, are leading the way. While there are clear reasons why these sectors are maturing at a rapid rate—e.g., the clear value proposition of helping to save lives following a disaster—there is a clear need, and opportunity, for public sector and civil society actors across other sectors to be more imaginative in considering how data from other sectors can help them achieve their missions and better address public problems—from improving health outcomes to more targeted urban planning and beyond.

III. SOCIAL MEDIA DATA COLLABORATIVES

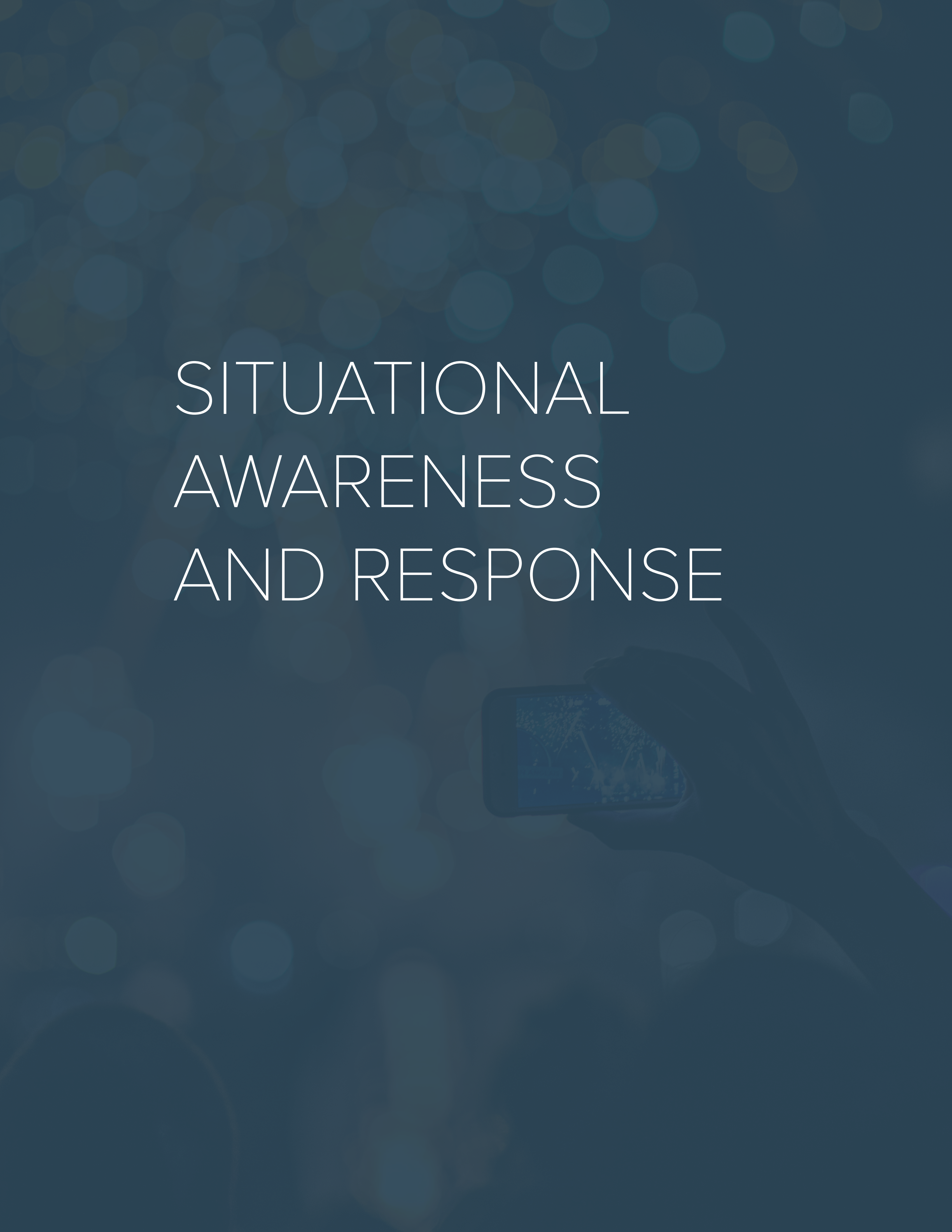
In Action

In what follows, we examine a number of existing social media data collaboratives that illustrate the five central value propositions introduced above. These initiatives were selected based on their current state of implementation, and diversity in terms of region, sector and partners. The case studies inform the key takeaways and recommendations at the conclusion of this report.

Intended Public Benefit	Case	Type of Data Collaborative	Sector	Region
Situational Awareness and Response	Facebook Disaster Maps	Trusted Intermediary	Crisis Response	Global
Situational Awareness and Response	Tracking Anti-Vaccination Sentiment in Eastern European Social Media Networks	API	Health	Eastern Europe
Situational Awareness and Response	Facebook Population Density Maps	Intelligence Product	Infrastructure	Global, focusing on developing countries
Knowledge Creation and Transfer	Yelp Dataset Challenge	Prizes and Challenges	Various	United States
Knowledge Creation and Transfer	MIT Laboratory for Social Machines' Electome Project	Intelligence Product	Politics and Elections	United States
Knowledge Creation and Transfer	LinkedIn's Economic Graph Research Program	Prizes and Challenges	Business	United States
Public Service Design and Delivery	Facebook Future of Business Survey	Trusted Intermediary	Business	Global
Public Service Design and Delivery	Easing Urban Congestion Using Waze Traffic Data	Trusted Intermediary	Infrastructure	Global
Public Service Design and Delivery	Facebook Insights for Impact Zika	Intelligence Product	Health	Brazil
Prediction and Forecasting	Tracking the Flu Using Tweets	API; Prizes and Challenges	Health	United States
Prediction and Forecasting	Predicting Floods with Flickr Metatags	API	Crisis Response	United Kingdom
Prediction and Forecasting	Predicting Adverse Drug Events by Mining Health Social Media Streams and Forums	API	Health	United States, Spain

Intended Public Benefit	Case	Type of Data Collaborative	Sector	Region
Impact Assessment and Evaluation	Sport England's This Girl Can	API	Health	United Kingdom
Impact Assessment and Evaluation	Using Twitter Data to Analyze Public Sentiment on Fuel Subsidy Policy Reform in El Salvador	API, Research Partnership	Energy	El Salvador
Impact Assessment and Evaluation	Measuring Global Engagement on Climate Change	API	Climate	Global

FIGURE 5



SITUATIONAL AWARENESS AND RESPONSE

Social media data collaboratives can improve situational awareness and response. Forming partnerships with social media companies can assist humanitarian organizations and other public sector entities in myriad ways—for example, by providing resources and analyses to which those actors would not otherwise have access. Such data collaboratives can deliver on humanitarian missions of improving public value, and can more efficiently provide visibility into the needs of vulnerable communities.

In this section, we examine three projects that focus on situational awareness and response. One case study **analyzes** public perceptions, and the other two leverage diverse datasets to map and **analyze** population movements and trends.

Case	Type of Data Collaborative	Sector	Region
Facebook Disaster Maps	Trusted Intermediary	Crisis Response	Global
Tracking Anti-Vaccination Sentiment in Eastern European Social Media Networks	API	Health	Easter Europe
Facebook Population Density Maps	Intelligence Product	Infrastructure	Global, focusing on developing countries

FIGURE 6

FACEBOOK DISASTER MAPS

Featured Statistic or Number: Over 7 million people marked themselves as safe using the Safety Check feature following the 2015 earthquake in Nepal.[25]
Type of Data Collaborative: Trusted Intermediary
Sector: Crisis Response
Region: Global

PROBLEM STATEMENT

As Mark Zuckerberg noted in his 2017 post on using Facebook to build a global community,[26] "Many dedicated people join global non-profit organizations to help, but the market often fails to fund or incentivize building the necessary infrastructure... There is a real opportunity to build global safety infrastructure."

In the face of humanitarian crises, namely natural disasters, international organizations and first responders are in need of useful data on where people are, where they are going, if they are safe, and where resources should be targeted. Unfortunately, as Toby Wicks, data strategist at UNICEF notes,[27] "Collecting high-quality data can be a time-consuming and costly exercise."

According to Arif Husain,[28] chief economist at the World Food Programme (WFP), the struggle to access and mobilize useful data is particularly challenging during the first 72 hours of a crisis. The difference between life and death following tsunamis, earthquakes and other natural disasters can amount to having visibility into the critical needs and areas where relief efforts should focus.

RESPONSE

In a collaboration with the International Federation of Red Cross and Red Crescent Societies (IFRC), UNICEF, and WFP, Facebook is attempting to fill that information gap with location and self-reported safety data collected from users in affected regions following natural disasters.[29] **Aggregated data**[30] on where people are, where they move over time, and where people have identified themselves as safe through Facebook's Safety Check feature is being shared with these humanitarian organizations in the wake of crises. Molly

Jackman, Facebook's Public Policy Research Manager,[31] believes that, "Facebook can help response organizations paint a more complete picture of where affected people are located so they can determine where resources—like food, water and medical supplies—are needed and where people are out of harm's way."

Unlike the more openly accessible Google Crisis Map,[32] the Facebook Disaster Maps initiative is intentionally built around collaborations between Facebook and its humanitarian organization partners, many of whom were engaged early in the development process to help steer efforts toward the greatest impact.

As Facebook's first data sharing initiative under its Data for Good program, the Facebook team knew it was balancing the need to provide organizations the information they needed to make accurate decisions during a natural disaster, with the need to protect user privacy and ensure that the data was shared in a responsible manner. Cooperation with trusted partners in the early stages allowed the program to take more risks to build datasets that the partners could realistically leverage during disasters.

Paige Maas, a Facebook data scientist behind the Disaster Maps initiative, saw "talking to experienced humanitarians" as a way to build an understanding of "where their data shortfalls were and how our data could potentially fill those gaps."[33] Early discussions also helped make clear what levels of temporal and spatial granularity would be most valuable in the shared data and maps.

To begin the collaboration in earnest, Facebook[34] demonstrated to IFRC how its data and data-driven maps could shine a light on three historical disasters: and the Kaikoura earthquake in 2016, the Fort McMurray Wildfire in 2016 and Hurricane Matthew in 2016. The demonstration showed how Facebook's data could provide insight into where and how people move following natural disasters.

Wicks[35] is hopeful that datasets provided by Facebook can help UNICEF "respond to critical needs and deliver relief in quicker, more agile ways that are uniquely tailored to the children and women we serve."

OUTCOME/DELIVERABLE/PRODUCT

Facebook's Disaster Maps initiative provides partner organizations with three types of maps[36] immediately following a natural disaster:

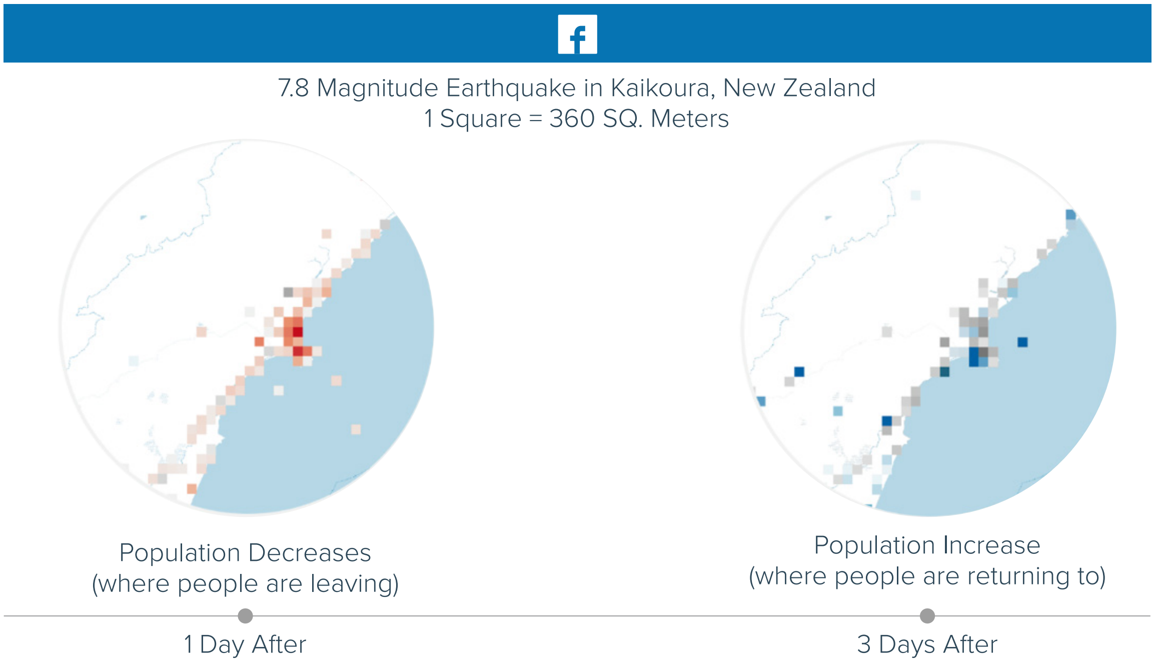
- *Location density maps* showing people's location (at an aggregate level) immediately before, during and after a natural disaster.
- *Movement maps* between neighborhoods or cities following a disaster, providing insight into common evacuation routes or potential barriers to movement.
- *Safety Check maps* showing where Facebook users have (and have not) used the platform's Safety Check feature to mark themselves as safe following the disaster.

Maas notes that the Disaster Maps initiative was developed with a keen focus on the question, "How can we make sure the data is useful while still protecting people's privacy?"[37] With that question in mind, each of the three types of map seek to strike a balance where data is granular enough to be useful to humanitarians, but aggregated to a level that obscures the release of any potentially harmful personally or demographically identifiable information. Whether or not Facebook and its partners are able to maintain this balance will only become clear over time.

Upon Disaster Maps' launch, Jono Anzalone, vice president of international services at American Red Cross, described how the types of data and maps on offer could benefit humanitarian organzations: "Maps help us in so many ways—from distributing relief supplies to preparing communities for disasters. By sharing anonymized location, movement, and Safety Check data with the American Red Cross, Facebook is helping us sharpen the essential tools we need for targeting communities in need, delivering aid, and fighting disease."[38]

IMPACT

The Disaster Maps project was officially launched just two months before the writing of this report, so it has, unsurprisingly, had little direct impact to date.


7.8 Magnitude Earthquake in Kaikoura, New Zealand
1 Square = 360 SQ. Meters
Population Decreases
(where people are leaving)
Population Increase
(where people are returning to)
1 Day After
3 Days After

This type of information can help response organizations understand which neighborhoods suffered the most damage following an earthquake and where people might be in need of help as they evacuate their homes and eventually return.

FIGURE 7

In a post[39] written by the engineers, researchers and other Facebook personnel behind Disaster Maps, they advocate for humanitarian organizations "to use our data as a part of a broader set of data that help inform resource deployment. Specifically, they should take into account that our data represents this specific population and to consider it in the context of the other information they receive."

While much of the use of Disaster Maps immediately following its launch aimed to verify if and how it could be useful through the analysis of previous disasters, the real-time use of the maps commenced in summer 2017. In its first two real-time uses, Disaster Maps were shared with regional actors following the flood disaster in Assam, India, and with the Red Cross and World Food Programme at the start of unprecedented rainfall and flooding in Houston during Tropical Storm Harvey and Hurricane Irma.[40]

The on-the-ground impacts of these efforts will not be clear for some time, but perhaps the biggest impact of Disaster Maps at this very early stage is the awareness building and inspiration it provides to other corporations and, in particular, social media companies. As Arif Husain argues, "Data is an indispensable resource in the world we live in, and we hope that other companies will come forward to support to bring their data, resources and expertise to support the world's efforts to achieve the SDGs."[41] Furthermore, Disaster Maps has laid the groundwork for understanding how to share data with organizations while maximizing the value of the data and the privacy of users by leveraging statistics and data processing to make the data safer and filter data signal from noise.

Toby Wicks from UNICEF similarly sees the Disaster Maps initiative as an early step toward broader data collaboration built on social media: "Given the rapidly changing landscape and our limited resources, the smartest approach is likely to be a heavy emphasis on demand-driven partnerships, rather than significant expansion of our own capacity. This is one great example of the direction we're going along that path."[42]

TRACKING ANTI-VACCINATION SENTIMENT IN EASTERN EUROPEAN SOCIAL MEDIA NETWORKS

Featured Statistic or Number: In 2012, 86 percent of anti-vaccine social media conversations occurred on blogs
Type of Data Collaborative: API
Sector: Health
Region: Easter Europe

PROBLEM STATEMENT

As is the case across many parts of the world, recent years have seen an increase in anti-vaccine sentiment among citizens in Central and Eastern Europe and the Commonwealth of Independent States. This skepticism—or outright hostility—toward vaccination among many parents has been attributed to a number of potential root causes, including complacency arising from high vaccination rates, and thus low levels of related illnesses; mistrust in institutions responsible for immunizations; concerns regarding supposed side effects like autism and HIV; and the increased availability of anti-vaccine messages across the Internet. Witnessing an uptick in fear and distrust surrounding vaccinations in Eastern Europe, UNICEF recognized that they and other actors responsible for children's health had little understanding of how anti-vaccine messages spread across the Internet—social media in particular—and what the key drivers of this hostility toward immunization were across different communities.

RESPONSE

To help UNICEF and other actors craft more effective pro-immunization messaging programs in the region, researchers at UNICEF set out to monitor social media networks, including blogging platforms, forums, Facebook, Twitter, Tumblr, and YouTube; analyze prevalent conversation themes according to volume, types of engagement, and demographics participating; identify influencers across languages and platforms; and develop specific recommendations for improving messaging strategies across languages, platforms and conversation themes.

The research methodology involved scraping text conversations from social media platforms in English, Russian, Polish and Romanian, cleaning and anonymizing the data, and categorizing those conversations toward identifying key patterns. The analysis covered the period from May 1st through July 30th in 2012.

OUTCOME/DELIVERABLE/PRODUCT

The key findings from the research effort were collected in a UNICEF working paper[43] released during European Immunization Week in April 2013. The report shared a number of key findings related to anti-vaccine sentiments and discussions, including:

- Blogs were the most frequently used channels for discussing anti-vaccination content in the identified languages—with 86 percent of relevant conversations in Romanian occurring on blogs, 85 percent of Polish conversations; 65 percent of those in Russian; and 47 percent of English conversations.
- Facebook hosted a quarter of anti-vaccine conversations occurring in English; and 24 percent of those occurring in Russian were on Twitter.
- Forums accounted for only a very small percentage of total anti-vaccination conversations (2 percent) but saw the most heavily engaged audience—with 25 percent of the total interactions among users (i.e., likes, comments, upvotes).
- Facebook saw a similar but slightly less pronounced disparity between total conversations and user engagement—with 9 percent of the total anti-vaccine conversations across languages occurring on Facebook along with 21 percent of all interactions.
- Future messaging efforts might also need to take into account gender differences that the working paper identified with respect to the content of posting on social media. For example, across all the languages studied, over 80% of those objecting to vaccines based on conspiracy theories, for example, were men, and 59% of those concerned about disabilities were women.

The UNICEF researchers also uncovered major differences regarding the drivers of anti-vaccine sentiment in different languages. Concerns over chemicals and toxins represented the vast majority in Romanian postings; side effects and chemicals and toxins were slightly more common than other issues in relevant postings from Polish speakers; nearly all anti-vaccine postings shared in Russian focused on religious and ethical beliefs; and a majority of posts in English tended to focus on either religious and ethical beliefs or conspiracy theories.

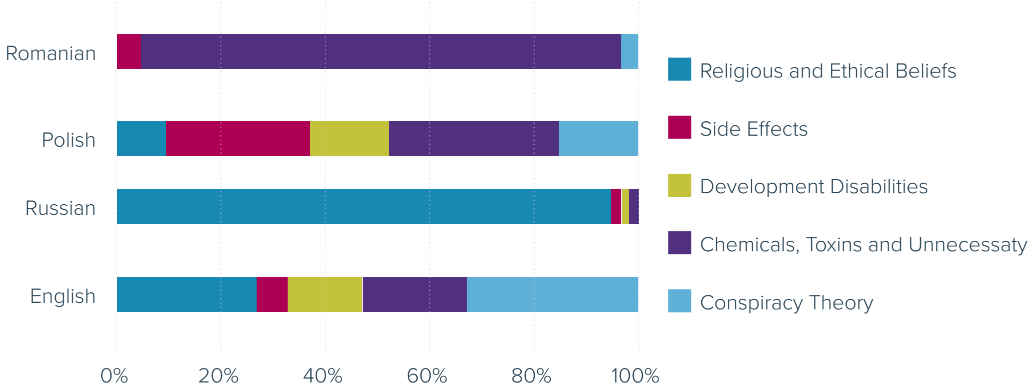


FIGURE 8: ALLOCATION OF ARGUMENTS BY LANGUAGE FOR THE ANTI-VACCIONATION SENTIMENT

IMPACT

To inform future communications efforts toward increasing vaccination rates in Eastern Europe, the researchers identified three key spheres inhabited by those participating in anti-vaccine conversations on social media:

- **Individual Sphere**—users driven by highly personal issues of concern or fear;
- **Contextual Sphere**—users driven by a distrust of governments, pharmaceutical companies, international organizations and other actors, without necessarily being directly hostile to vaccines; and
- **Transcendental Sphere**—users driven by more dogmatic negative attitudes toward vaccines derived from idealistic, religious or ethical beliefs.

This understanding helped UNICEF to develop targeted strategies for meaningfully engaging those inhabiting each sphere. Inhabitants of the Individual Sphere, for example, should be provided with pro-vaccine information and questions to ask doctors toward building confidence in the decision-making process. Public health advocates should target those in the Contextual Sphere, on the other hand, with hard facts on vaccines and focus on transparency related to testing, ingredients, potential side effects and other issues that can drive distrust. Finally, those in the Transcendental Sphere should be monitored but not directly engaged as their beliefs leave them disinterested in dialogue with pro-vaccine actors.

It is unclear to what extent these recommendations were taken up in practice by UNICEF, and, if those efforts occurred, what their impacts were on vaccination rates in Eastern Europe. The effort did inspire follow-on work in Indonesia[44] where the Ministry of Development Planning (Bappenas), the Ministry of Health, UNICEF, WHO, UN Global Pulse and Pulse Lab Jakarta sought to build a deeper understanding of drivers of anti-vaccination sentiment, focusing specifically on Twitter.

FACEBOOK POPULATION DENSITY MAPS

Featured Statistic or Number: Square kilometers analyzed to uncover population density: 21.6 million km
Type of Data Collaborative: Intelligence Product
Sector: Infrastructure
Region: Global, focusing on developing countries

PROBLEM STATEMENT

There are a number of potential strategies being pursued to connect the billions of people around the world who still lack Internet access. Cellular technologies are broadly beneficial across the globe, but Wi-Fi hotspots and other limited-range access networks may prove useful for communities where people are separated by shorter distances, while cellular technologies are more beneficial for all people across the globe. Backhaul networks also have the potential to create internet links, placing users closer together through a string of terrestrial point-to-point connections or via an aerial backhaul solution, such as unmanned aerial vehicles or satellites, for especially scattered settlements.[45]

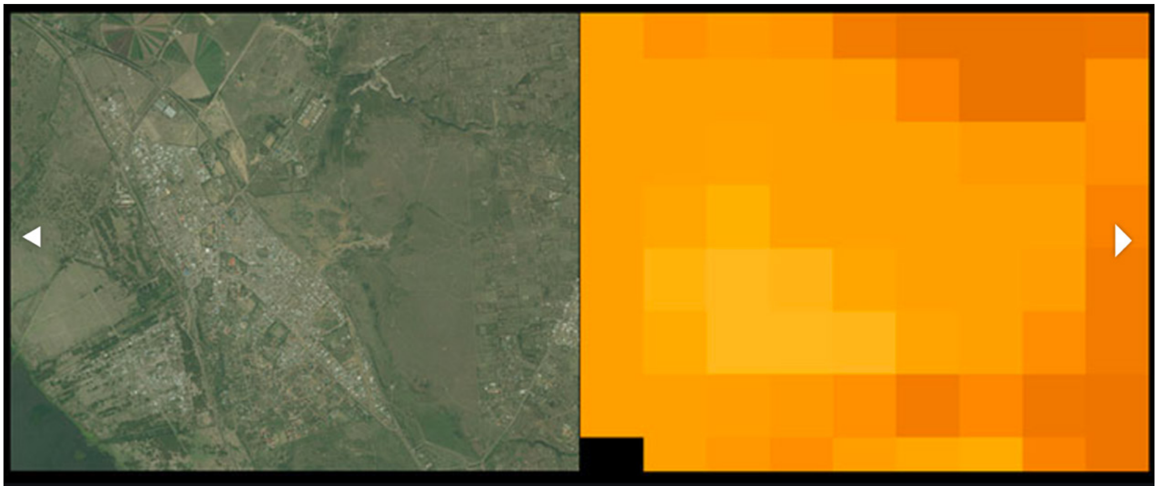
However, we have little insight into which technological solution should be used in which locations, because we lack accurate knowledge about the population distribution in disconnected regions across the developing world. Moreover, creating a dataset with high spatial resolution could be a large undertaking for the countries that could benefit most from better Internet connectivity. Among other issues, this data gap creates challenges for researchers and international organizations faced with the challenge of designing the right connectivity infrastructure to provide a more granular insight on population distribution, considering aggregate population counts and civil censuses are insufficient when standing alone.

RESPONSE

To gain a better grasp of population density in disconnected parts of the world, the Facebook Connectivity Team, along with its Core Data Science, Infrastructure and FAIR and Applied Machine Learning teams[46] sought to collaborate with a number of actors from across sectors. Working with data from DigitalGlobe, a satellite imagery company, focusing on 20 countries from across the developing world, the Facebook Connectivity Team reconfigured Facebook's image-recognition engine to detect whether a satellite image contained a building. From there, the team developed a neutral learning network that could automatically identify outlines of buildings with high confidence.

Following the initial creation of the dataset through Facebook's collaboration with DigitalGlobe, Facebook engaged Columbia University and the World Bank[47] to further improve the quality of its data, and to explore practical uses—related to Internet connectivity and beyond. The team from Columbia used census data to validate the Facebook population density datasets and enable comparison and works toward more "coarse-grained datasets." The World Bank's Living Standards Measurement Study (LSMS) program then further validated these improved datasets through anonymized household surveys undertaken in the priority locations.

With the help of the collaborators, Facebook then used machine-learning AI to handle large datasets and provide quick analyses of targeted countries in a turn-around time of just two weeks. From the resulting data, Facebook was able to develop internal tools that could test whether approaches were effective in a matter of hours.



DigitalGlobe satellite image of Naivasha, Kenya (left) and gridded population of the world v4 from CIESIN at Columbia University (right).

FIGURE 9

OUTCOME/DELIVERABLE/PRODUCT

The Population Density initiative has yielded a staggering 350 TB of imagery[48] across the 20 priority developing countries. The datasets can be accessed and downloaded at the Columbia University Center for International Earth Science Information Network (CIESIN) website.[49]

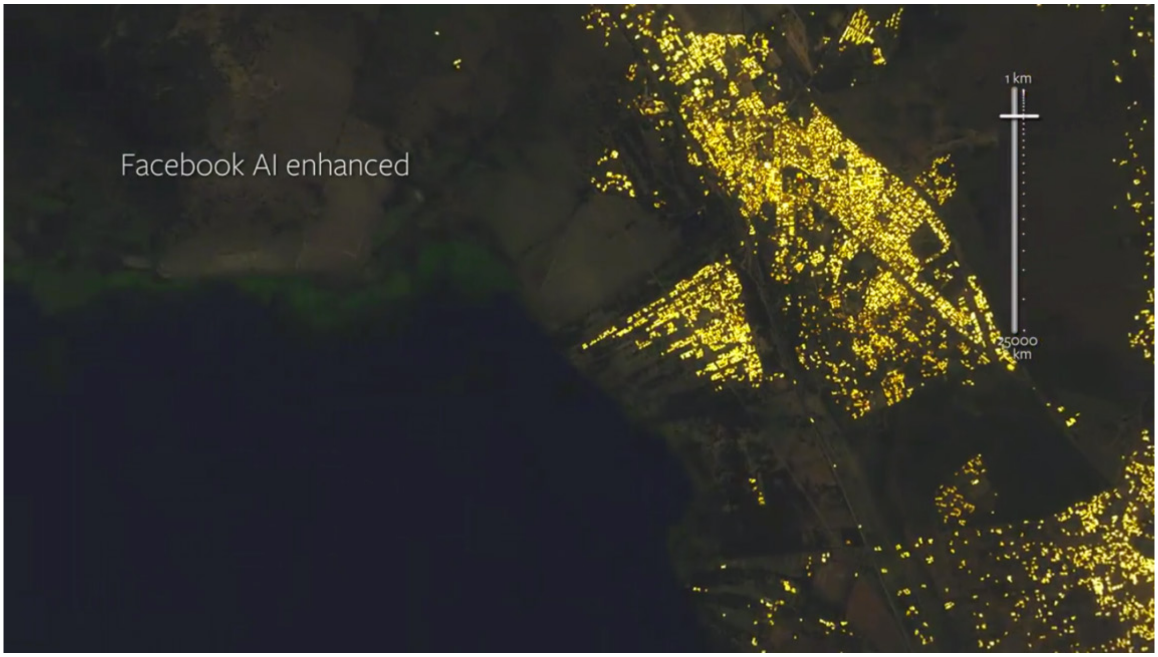


FIGURE 10

IMPACT

In total, the collaboration enabled the analysis of 21.6 million km^2 of the priority countries, based on the processing of 14.6 billion images through the Facebook neural network—**more than ten times as much as all the images analyzed by Facebook on a daily basis**.[50] The dataset generated "provides the most accurate estimates of population distribution and settlements"[51] for each target country. The settlement and population maps built from this data focus down to the level of 5 meters, and are the most granular of any dataset currently in existence.[52]

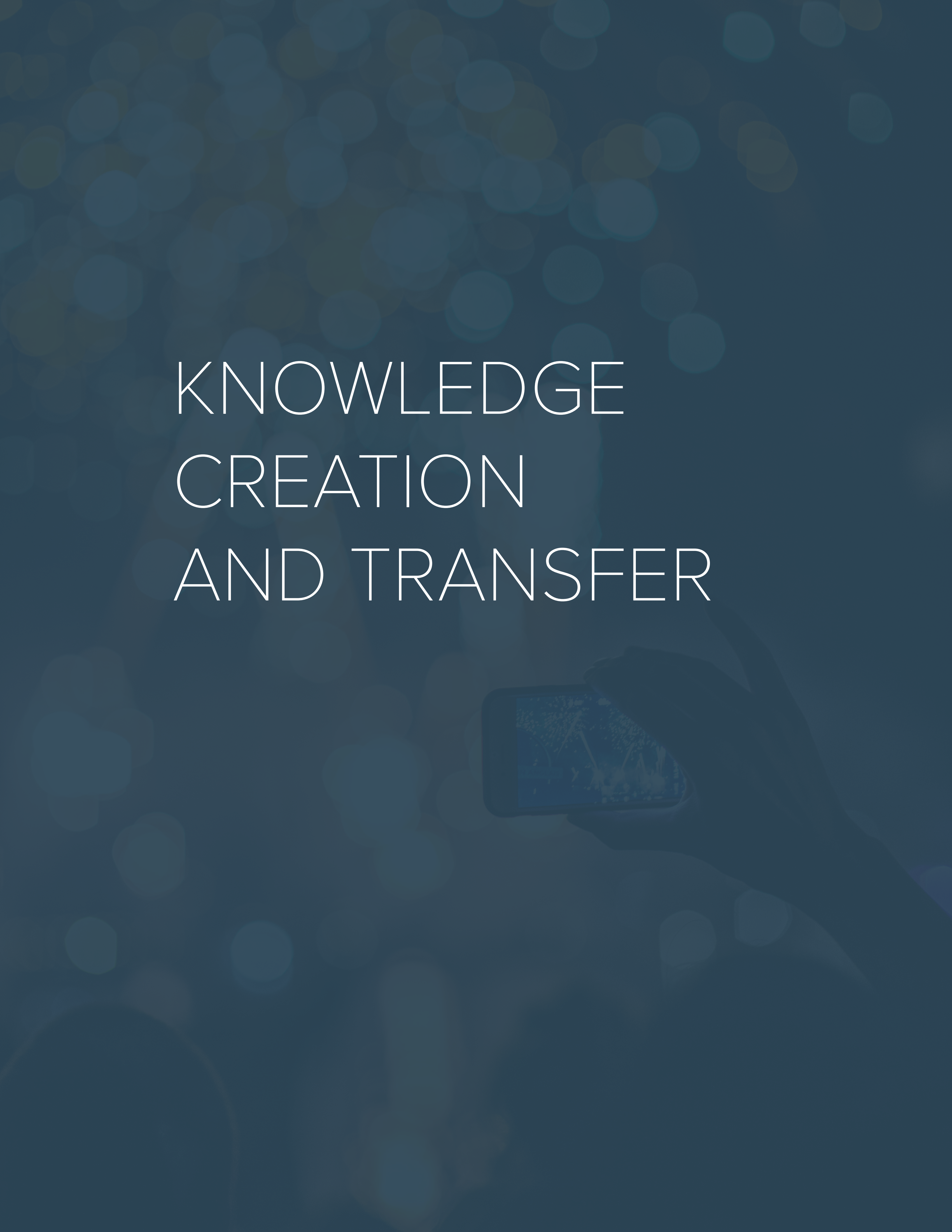
As of August 2017, the Facebook Connectivity team has not publicly shared what impacts these new insights into population density have had on its efforts to connect more people across the developing world to the Internet.

Robert Chen, the director of CIESIN argues that, "The innovative population distribution data produced through this collaboration will not only be useful to Facebook, governments, and other public and private groups in optimizing efforts to expand Internet access, but will also be available to the broader development community as a unique resource to help all parts of the world achieve the Sustainable Development Goals, wherever people live."[53]

To date, the Red Cross and Missing Maps program used the population density maps in order to inform a measles and rubella campaign in Malawi in partnership with the Malawi Ministry of Health and other international NGOs. Three thousand local Malawi Red Cross volunteers in these districts were trained on key messages and conducted roughly 100,000 house-to-house visits over three days. Volunteers ensured that the households heard about the vaccination campaign and provided an opportunity to ask questions about the vaccine and the campaign – providing an important avenue to address the concerns of those who might not otherwise have brought their children to be immunized. Facebook's high-resolution population data for Malawi, helped the Red Cross and Missing Maps in identifying areas with and without concentrations of people. This made the tracing much, much faster allowing the Missing Maps to filtered 97% of land space that were uninhabited areas out of the tracing tasks so that they could direct remote mappers towards areas that had features for them to map and thus get volunteers to households more efficiently.[54]

Specialists at the World Bank[55] are also "considering the use of the Facebook population maps for specific projects in infrastructure planning and impact assessment, and in disaster relief activity planning and scenario analysis," but it is not yet clear if those efforts have begun in earnest.

KNOWLEDGE CREATION AND TRANSFER



In the following section, we examine three case studies: one that guides media-covered conversation to publicly trending topics, and two that use social media data to enable prizes and challenges for uncovering new data-driven insights to benefit the public good.

Case	Type of Data Collaborative	Sector	Region
Yelp Dataset Challenge	Prizes and Challenges	Various	United States
MIT Laboratory for Social Machines' Electome Project	Intelligence Product	Politics and Elections	United States
LinkedIn's Economic Graph Research Program	Prizes and Challenges	Business	United States

FIGURE 11

YELP DATASET CHALLENGE

Featured Statistic or Number: $65,000 awarded to 13 different teams of students over four years
Type of Data Collaborative: Prizes and Challenges
Sector: Various
Region: United States

PROBLEM STATEMENT

Researchers and problem-solvers often lack access to aggregate-level data on aspects of the economy and built environment that could help to uncover new insights into public life. Due to this lack of accessible information, they are often unable to conduct targeted data analysis of the otherwise public datasets.

RESPONSE

In 2013 Yelp set out to organize a data challenge, providing public access to its crowd-sourced review and ratings data in a way that could be tightly controlled and orchestrated by the company.[56] Cash rewards were offered to winning teams of students submitting original research that focused on using the data in innovative and analytical ways through myriad categories—including cultural trends, seasonal trends, and business **changepoints**.

To allay privacy and competitive advantage concerns, Yelp took precautions **to protect their** shared data. Chief among them was the limited amount of data offered to research teams. Although the data was made publically available on the Yelp website for anyone to download, the dataset offered in the first round only featured reviews and ratings from the Phoenix metropolitan area. Moreover, the company emphasized that only academic, non-commercial applications of the data would be permissible.[57]

OUTCOME/DELIVERABLE/PRODUCT

The first round of the Yelp Dataset Challenge was announced in March 2013 and invited any student to submit an inventive insight or analysis built from the provided dataset.[58] Winning teams of students, as chosen by Yelp's data mining engineers, received $5,000 and the potential to win another $1,500 for publishing their work in a paper and presenting at a conference.[59] In recent years, Yelp has greatly expanded the dataset it provides for every challenge. The dataset from the ninth round, announced in January 2017, featured ratings and reviews from 11 different cities in four different countries in addition to over 200,000 images of the included businesses.[60]

In the four years since the beginning of the challenge, Yelp has awarded $65,000 to 13 different teams of students. Notable past projects include ***On the Efficiency of Social Recommender Networks***[61] from Princeton University and ***Personalizing Yelp Star Ratings: A Semantic Topic Modeling Approach***[62] from Yale University. More recent winning projects include ***From Group to Individual Labels Using Deep Features***[63] by a team at University of Oxford and University of California Irvine, and ***Semantic Scan: Detecting Subtle, Spatially Localized Events in Text Streams***[64] by a team from Carnegie Mellon University and University of Notre Dame.

These projects and reports are based on only a very small fraction of the data Yelp has collected across the world. Yelp will not honor requests from students for datasets outside the scope of datasets currently available. Still, though individuals who are not currently students are not able to participate in the challenge itself, they are free to use the datasets for themselves for academic, non-commercial purposes.

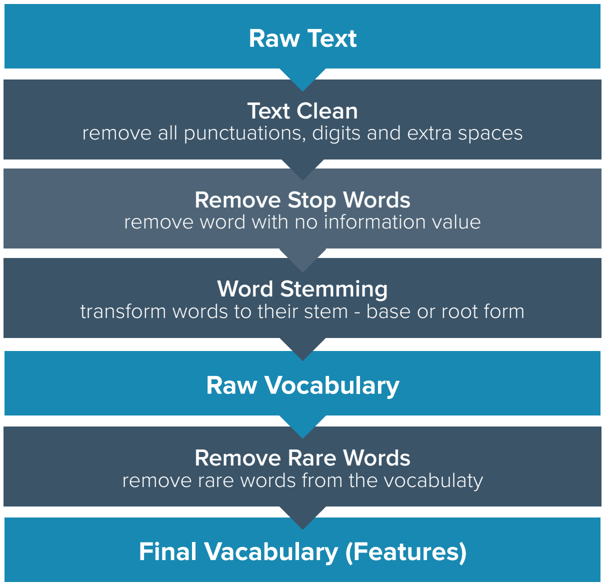

Raw Text
Text Clean
remove all punctuations, digits and extra spaces
Remove Stop Words
remove word with no information value
Word Stemming
transform words to their stem - base or root form
Raw Vocabulary
Remove Rare Words
remove rare words from the vocabulaty
Final Vocabulary (Features)

FIGURE 12: DATA PREPROCESS FLOW

Source: ***Prediction of Yelp Review Star Rating Using Sentiment Analysis***, pg. 2

IMPACT

Machine learning and natural language processing classrooms around the United States have incorporated the dataset challenge into their curriculums as course projects and assignments.[65,66] Hundreds of papers have been published citing the Yelp dataset as a source and many more teams of students have made a submission to the challenge.[67]

As an example of the type of knowledge being created and publicly shared, one of the first winners and the most widely cited paper resulting from the challenge came from a team analyzing text reviews in order to improve the intelligence of recommender systems. Recommender systems are ubiquitous forms of artificial intelligence playing a key role in the data-driven economy. The team found that a recommender system—like Amazon product recommendations or Spotify playlist suggestions, among others—could be brought up to speed for new users more quickly by taking the text of a review into account in addition to its rating.[68]

Another winning team devised a way of analyzing reviews to determine separate ratings for hypothetical subtopics (such as service, decor or food quality) allowing consumers to make better-informed decisions based on their preferences and business owners to understand what aspects of their business are most in need of improvement.[69]

MIT LABORATORY FOR SOCIAL MACHINES' ELECTOME PROJECT

Featured Statistic or Number: The day before the 2016 U.S. election, 43% of relevant Twitter conversations focused on foreign policy and national security; 15% on immigration; and 13% on guns.

Type of Data Collaborative: Intelligence Product

Sector: Politics and Elections

Region: United States

PROBLEM STATEMENT

Until very recently, the public had limited options to engage meaningfully with political candidates during campaigns.[70] Chatter of limited importance tends to command news coverage; horse race reporting is an often criticized, but seemingly inevitable staple of political media coverage.[71] The media and the candidates also drive the broad-reaching conversation, while citizens and individuals often turn to social media echo chambers to discuss their political beliefs. Twitter in particular has become a favourite among politicians, journalists and everyday citizens alike for opining on the political news of the day.[72,73] The sheer number of tweets— hundreds of thousands per day—presents a considerable challenge to traditional methods of analysis and reporting on this chatter.[74] How can this online collection of the real-time reactions and opinions of everyday Americans bring further insights and analysis? Is there a way to analyze data and open interactions on Twitter that is representative of public opinion on political matters?

RESPONSE

The Laboratory for Social Machines at MIT's Media Lab is well equipped to study the behavior of political conversations on social media due to its expertise in mapping and analyzing social systems using data science methods, and the lab's access to Twitter's firehose of real-time and historical tweets thanks to a gift from the social media company.[75,76]

In order to begin a pilot study of political conversations on Twitter, researchers developed two stages of machine learning algorithms. The first stage of algorithms learned to identify those tweets that were related to the election. This body of generally relevant tweets was then passed to a second stage of algorithms which sorted the tweets by topic based on the specific issue being discussed. To supplement their study, the researchers also performed this analysis on thousands of news articles harvested every day from over 30 news sites.[77]

In the end the researchers were successful in establishing a corpus of 2016 election-related tweets and news articles sorted according to 21 different policy issues including national security, the economy, healthcare, education, immigration, taxes and so on. As new material was classified and added to the corpus every day, the researchers were able to track how the magnitude of conversation related to any one of these issues changed over time.

OUTCOME/DELIVERABLE/PRODUCT

The centerpiece product of these analyses became the interactive visualizations [see image below], available on the Electome website for anyone to see and manipulate, depicting which issues and topics sparked the greatest share of the election-related conversation over the course of the 2016 presidential election.[78] These data visualizations are primarily available in two forms. For each of the debates, conversations and topics can be tracked minute by minute.[79] For the election season as a whole, the most active issues can be tracked on a daily basis.[80] Also, the results of the Electome project are graphical visualizations depicting how Twitter users who regularly engage in election-related conversation are connected to other similarly engaged Twitter users.

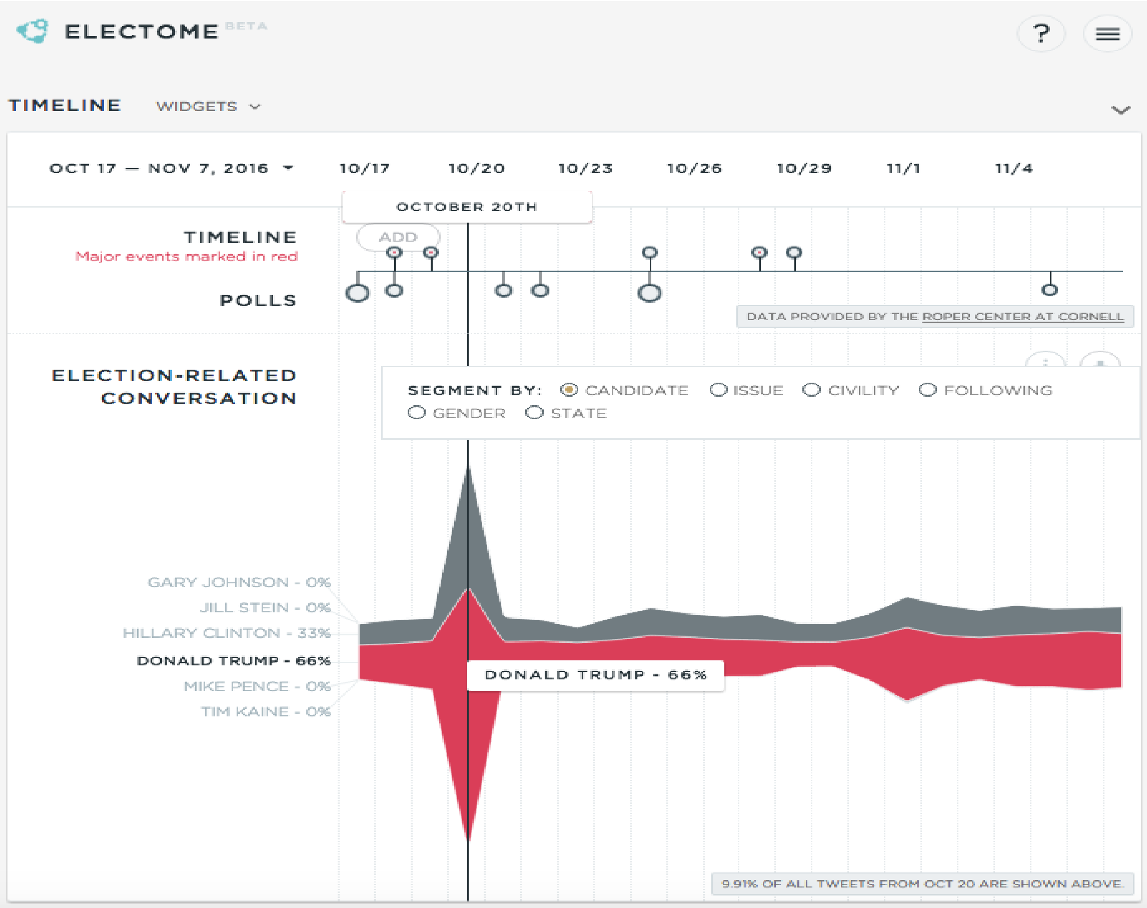
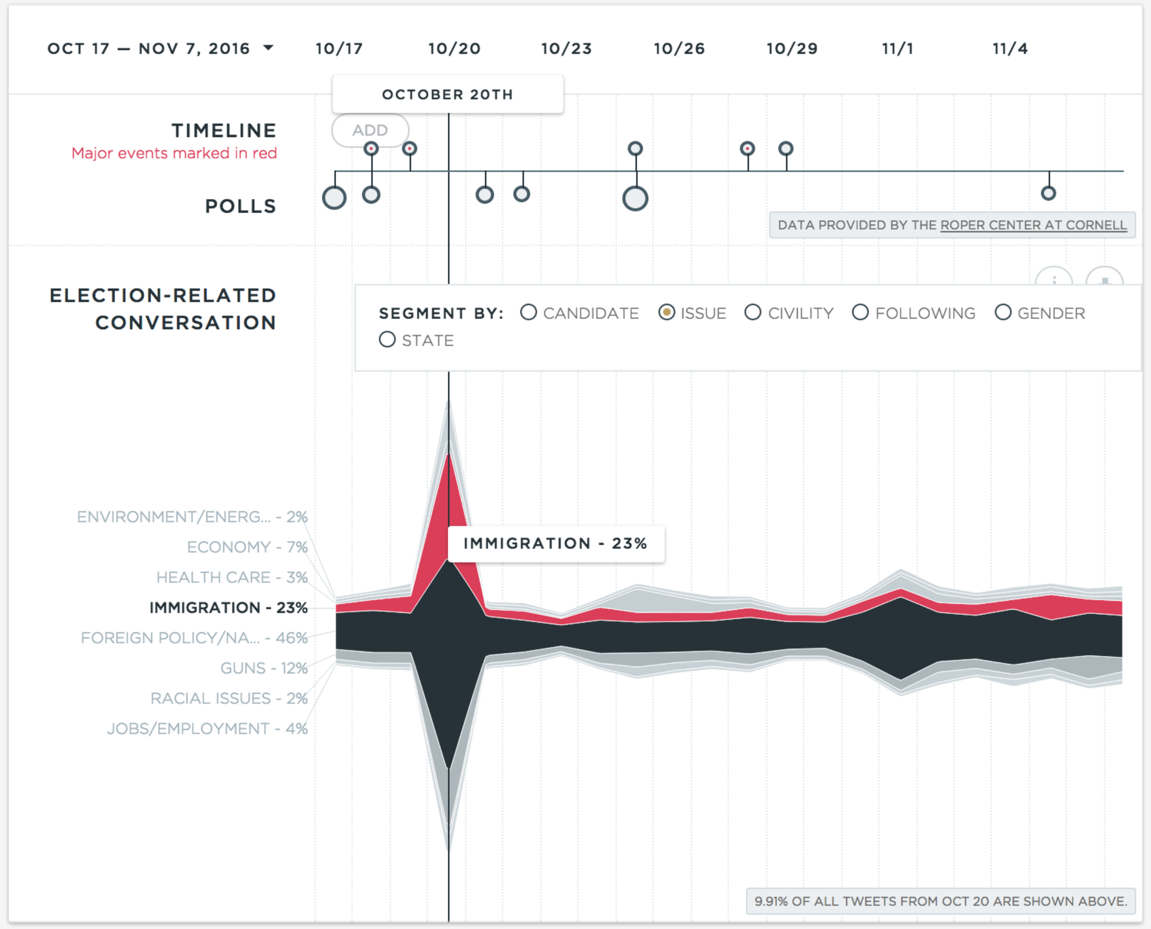


FIGURE 13[81]

IMPACT

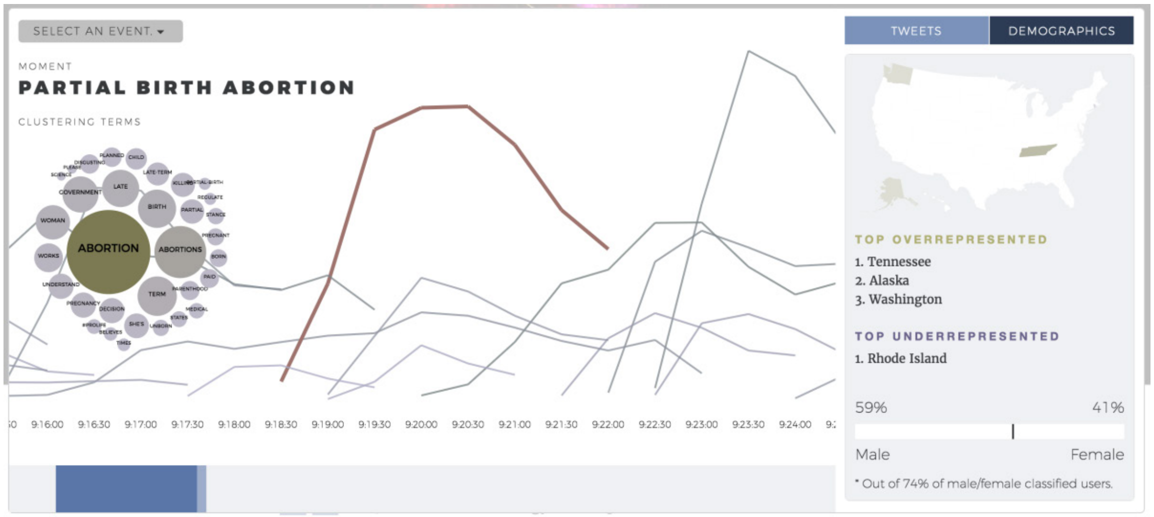
This first set of visualizations, depicting how interest in various issues on Twitter waxed and waned throughout the election season, is impactful first and foremost in providing a new body of political knowledge focused on the issues rather than metrics for winning or losing.



SEGMENT BY: ○ CANDIDATE ● ISSUE ○ CIVILITY ○ FOLLOWING ○ GENDER ○ STATE

9.91% OF ALL TWEETS FROM OCT 20 ARE SHOWN ABOVE.

FIGURE 14



MOMENT

PARTIAL BIRTH ABORTION

TOP OVERREPRESENTED

1. Tennessee
2. Alaska
3. Washington

TOP UNDERREPRESENTED

1. Rhode Island

59% Male

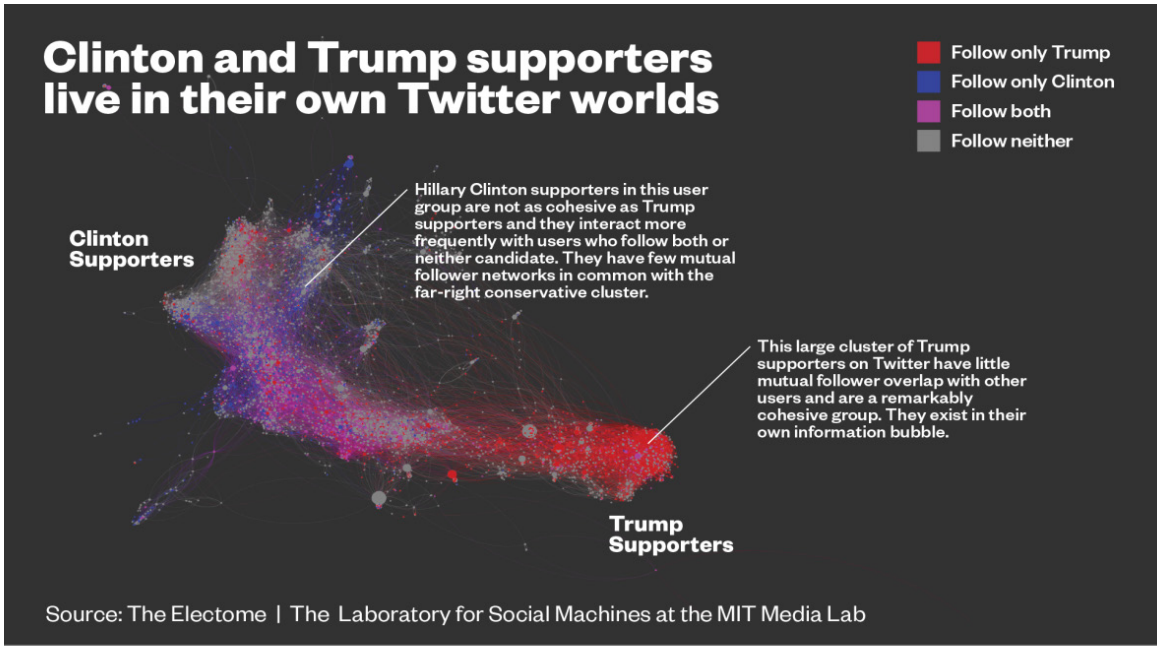
41% Female

* Out of 74% of male/female classified users.

FIGURE 15

In this respect, the Electome serves as a real-time tool to help news sites to be more responsive to the issues that matter the most to the tweeting public. For example, early Electome results revealed that the media was spending a disproportionate amount of coverage on campaign finance, though Twitter conversations were dominated by national security and immigration. The Electome provided their analysis to various news sites such as the Washington Post,[82] Bloomberg News,[83] The Wall Street Journal,[84] Fusion,[85] Vice News[86] and CNN[87] to inform coverage of the election.[88] The Electome also partnered with the Commission for Presidential Debates to provide analysis to the presidential debate moderators.[89]

The Electome visualizations also shed light on other phenomena such as increased political polarization. The mapping of active political tweeters and their connections reveals how Clinton and Trump supporters were clustered in two isolated groups. Mainstream reporters overlap greatly with the former group and hardly at all with the later group, offering support for the idea of political bubbles.[90]

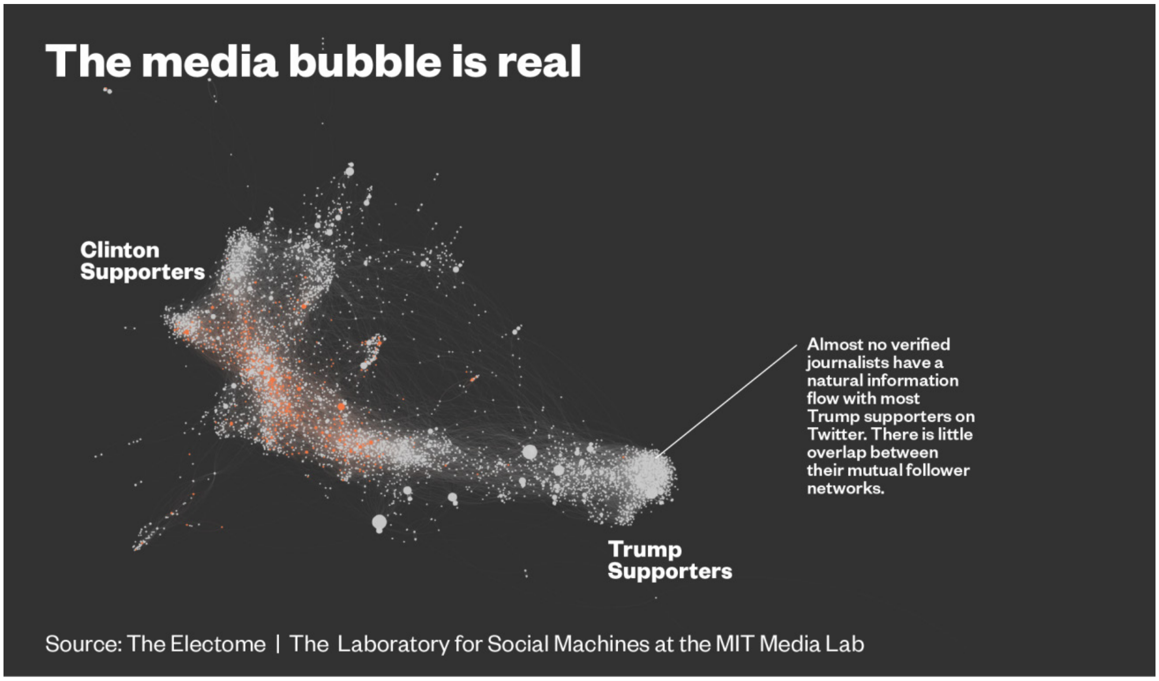


Hillary Clinton supporters in this user group are not as cohesive as Trump supporters and they interact more frequently with users who follow both or neither candidate. They have few mutual follower networks in common with the far-right conservative cluster.

This large cluster of Trump supporters on Twitter have little mutual follower overlap with other users and are a remarkably cohesive group. They exist in their own information bubble.

Source: The Electome | The Laboratory for Social Machines at the MIT Media Lab

FIGURE 16



Source: The Electome | The Laboratory for Social Machines at the MIT Media Lab

FIGURE 17

Nevertheless, the Electome is a limited resource. First and foremost, the Electome can only claim to represent the members of the public who actually tweet, and only about 21 percent of American adults use Twitter[91]—with even fewer users being active. An article in the ***Washington Post*** highlighted significant differences in a public opinion poll administered to the public at large and the analytics offered by the Electome.[92] This discrepancy calls into question the usefulness of the Electome's Twitter analysis in isolation.

LINKEDIN ECONOMIC GRAPH RESEARCH PROGRAM

Featured Statistic or Number: 11 finalists selected to manipulate LinkedIn's data to create useful economic insights
Type of Data Collaborative: Prizes and Challenges
Sector: Business
Region: United States

PROBLEM STATEMENT

Researchers often lack access to useful aggregate information related to different aspects of the economy that, if available, could help uncover useful insights and potentially spur job growth and economic activity. While there exists some data on the types of jobs available, what skills are needed, what skills individuals looking for jobs possess, more granularity and diversity of datasets on these and related topics could greatly enhance economic research efforts.

RESPONSE

LinkedIn sought to increase access and use of such economic data starting in 2014 with the Economic Graph Challenge.[93] The challenge asked teams of academics from universities, think tanks, NGOs and other non-profit institutions to submit proposals for research projects using LinkedIn data. The winning teams would be given $25,000 and the opportunity to implement their proposals as year long research projects using LinkedIn data. Unlike Yelp's challenge, LinkedIn's data challenge invited researchers of all experience levels to submit proposals, but only made the data available to those proposals that were accepted.

LinkedIn's concerns about data security are evident in the design of the challenge and research program that they established. The researchers selected to work with LinkedIn are given limited access to the data. The data is only available in a sandbox environment and the data cannot be used beyond the scope of the specific research project they are working on. In addition, each researcher must undergo LinkedIn's data privacy training, and researchers must receive permission from LinkedIn before presenting or sharing any of the data they have analyzed. Finally, LinkedIn gives preference during the selection stages of the program to those proposals that would only require public LinkedIn data in aggregate form.[94]

OUTCOME/DELIVERABLE/PRODUCT

In the end LinkedIn received many more qualified proposals than it had expected. Instead of three finalists, LinkedIn allowed 11 teams to have access to its data for a year long research project followed by presentation at LinkedIn's headquarters.[95]

Encouraged by the attention garnered at its first attempt to opening up its data to outside academic research, LinkedIn decided to bring back the program in a slightly different form. In 2017 LinkedIn engineering announced the newly established Economic Graph Research Program, a program that would also solicit proposals from academics but with added emphasis on forming long-term partnerships with those researchers as opposed to the one-time competition model with hard deadlines.[96] Research proposals were in the process of being solicited and evaluated by LinkedIn as of summer 2017.

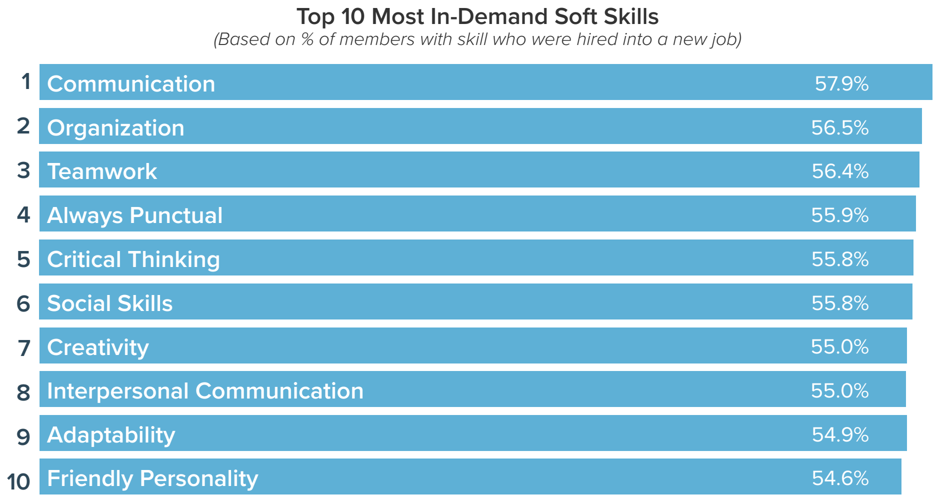


FIGURE 18

In tandem with its new research program, LinkedIn's in-house Economic Graph research team of economists and data scientists also compiles and publishes regular reports of its own analyses of LinkedIn data.[97]

IMPACT

Thanks to the 11 finalists originally selected, novel insights have been gleaned from the LinkedIn data made available for academic research. One group of researchers discovered that there were significant gender disparities in how MBAs promoted themselves on their LinkedIn profiles. A second group identified various micro-industries and examined how workers moved within or between them. A third group set out to measure human capital and to rank industries and companies by this metric.[98]

LinkedIn's own Economic Graph team also shares its research and insights into the professional networking data it has at its disposal. This Economic Graph team has provided various news sites with data and analysis to help inform their reporting.[99] Collaboration with the ***New York Times*** resulted in an article on the chances of successful promotion to the position of CEO and how they differ depending on a candidate's age, gender or previous experience.[100] Collaboration with the ***Wall Street Journal*** resulted in an article on the most in demand "soft skills"—e.g., communication, organization, teamwork—of that year.[101] Next, LinkedIn plans to expand the economic graph initiative to help address the skills gaps in the United States.

Social media data collaboratives can provide public sector entities with new evidence to inform the design and delivery of myriad public services. As evidenced through myriad initiatives around the world,[102] user-generated content on social media platforms can be analyzed toward better understanding public opinion and needs. New insights gained through such analyses can also inform transformative new approaches in infrastructure planning, sanitation efforts, and other areas of public concern.

PUBLIC SERVICE DESIGN AND DELIVERY

In the below section, we examine three initiatives leveraging social media data collaboratives to improve public service design and delivery. One case is focused on effective urban planning, while the other two aim to create new evidence-based interventions in the heath and economic sectors.

Case	Type of Data Collaborative	Sector	Region
Facebook Future of Business Survey	Trusted Intermediary	Business	Global
Easing Urban Congestion Using Waze Traffic Data	Trusted Intermediary	Infrastructure	Global
Facebook Insights for Impact Zika	Intelligence Product	Health	Brazil

FIGURE 19

FACEBOOK FUTURE OF BUSINESS SURVEY

Featured Statistic or Number: 90,000 businesses responded to survey in first seven months
Type of Data Collaborative: Trusted Intermediary
Sector: Business
Region: Global

PROBLEM STATEMENT

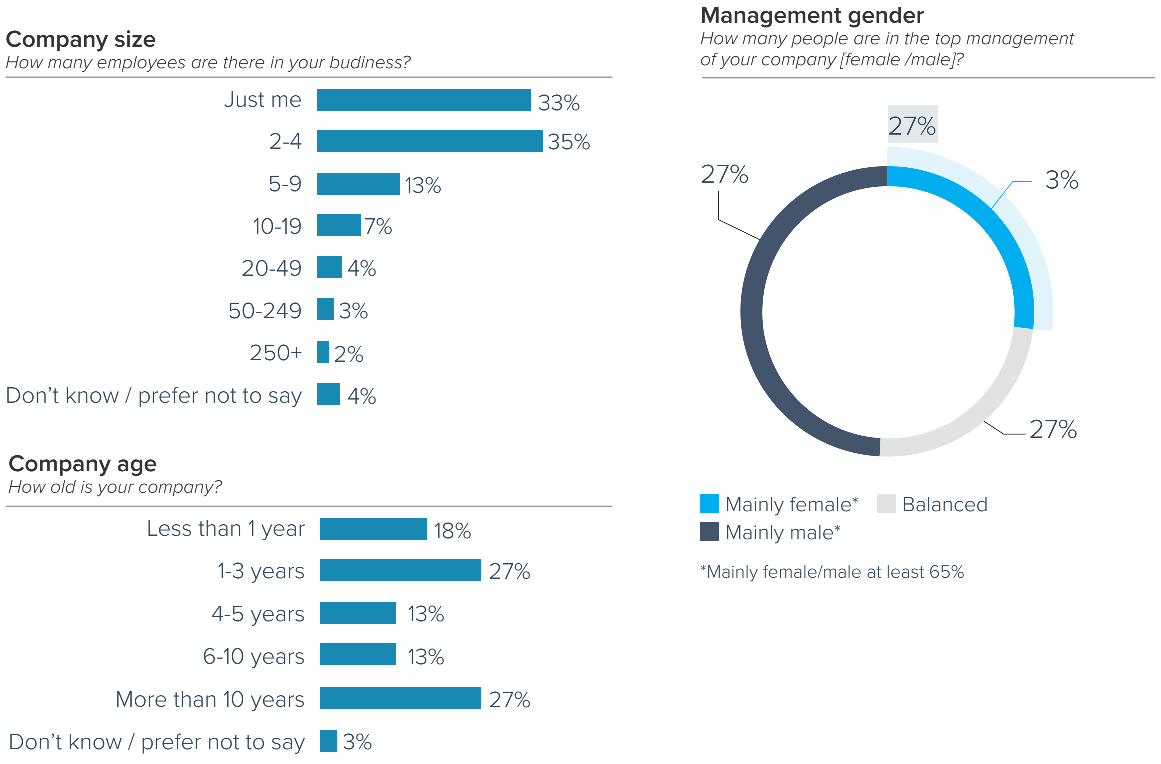
Fostering the success of businesses in a mobile economy demands a good understanding of their needs and environment. Small- and medium-sized enterprises (SMEs) are of particular interest to policymakers, researchers, and other businesses because they are often an important source of innovation and economic growth.[103] Getting accurate and timely information on SMEs is challenging even in developed countries with good government data collection because of the costs in time and money of administering traditional surveys and the difficulty of reaching SMEs. In developing countries, it has been all but impossible.

RESPONSE

Surveys conducted through social media offer a potent way to connect with business owners to gather such information and acquire a more complete picture of what is happening in those economies. As members of the Global Partnership for Sustainable Development Data, Facebook, the World Bank, and OECD developed the Future of Business Survey to provide timely, policy-relevant aggregated data to help inform decision-making at all levels.

OUTCOME/DELIVERABLE/PRODUCT

In February 2016, in collaboration with the World Bank and OECD, Facebook launched a 15 item monthly survey targeted at SMEs with an active Facebook business page from 13 high income countries and **nine** middle income countries.[104] The survey asked SMEs about current and future business activity; perceived challenges such as attracting customers, maintaining profitability, and economic uncertainty; business characteristics like industry type, company age, and number and gender of employees; and strategy questions about use of online tools and engagement in international trade. Two questions—about the current state of their business and their expectations for the next six months—are used to create a composite Business Confidence Indicator.[105]



*Mainly female/male at least 65%

FIGURE 20

The initial results were collated in the Future of Business Survey, which was released in September 2016. Subsequent reports have focused on management gender (January 2017), trade (January 2017), learning and growing together (May 2017), and international trade (August 2017).[106] A dashboard (www.futureofbusinesssurvey.org) offers tools to visualize and interpret the data by country and time, methodology, and anonymized data for download.

IMPACT

In the first seven months, the survey generated responses from over 90,000 businesses in 22 countries: US, Canada, Germany, Spain, France, UK, Poland, Brazil, Mexico, Argentina, Colombia, India, Australia, Indonesia, Thailand, Vietnam, Ireland, South Africa, Israel, Egypt, and Japan. Since the launch of the report, the survey has expanded to include more than 140,000 respondents in 42 countries,[107] and the organizers hope to scale globally to include other countries with a sufficiently high number of SMEs.

Business Confidence Indicator

FIGURE 5

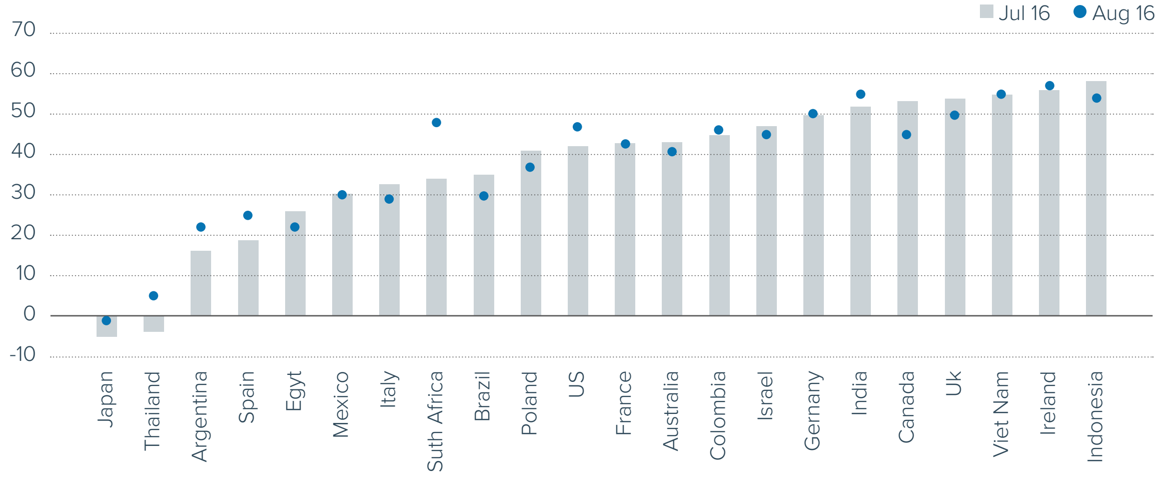


FIGURE 21

The Future of Business Survey is relatively new, making it difficult to point to specific instances of impact or influence. By filling data gaps with standardized data across countries, the organizers hope that going forward, the Survey will inform the market about SMEs and show how the technology industry and governments can support SME success.[108]

EASING URBAN CONGESTION USING WAZE TRAFFIC DATA

Featured Statistic or Number: Led to an 18 percent reduction in congestion in partner district in Boston
Type of Data Collaborative: Trusted Intermediary
Sector: Infrastructure
Region: Global

PROBLEM STATEMENT

Major cities throughout the world struggle with traffic congestion, creating delays and wasted driver and passenger time; fuel wastage and increased vehicle wear-and-tear; increased air and carbon pollution; stress, fatigue, road rage and diminished driver health; and increased collision rates. Cities with high rates of car use where infrastructure is challenged by existing, often antiquated construction or inadequate investment are particularly vulnerable to congestion problems. At the same time, real-time navigation apps are creating additional challenges for transport planning by encouraging drivers to take shortcuts through residential neighborhoods to shorten their commute times.

RESPONSE

Waze, the most widely used crowdsourced traffic and navigation platform, partnered with cities and government agencies to share publicly-available incident and road closure data through its free Connected Citizens Program. The collaboration was launched in 2014 with ten city partners, and now has over a hundred partners, including city, state and national government agencies, nonprofit organizations, and emergency services. Waze provides real-time incident and slow-down information sourced from drivers themselves, while partners contribute real-time information about construction, crash and road closures.[109]

OUTCOME/DELIVERABLE/PRODUCT

Partners have access to a portal, through which they can define the content and format of data they will share with Waze, and the content and data they will receive. Waze offers three datasets: Traffic View tool, which shows real-time traffic compared to historical data for the area; unusual traffic emails; and Waze XML/JSON API. In return, partners can share data about closures and incidents; crisis centers such as emergency shelters; speed limits, school zones, HOV lanes and license plate restrictions; real-time snow plow and garbage truck feeds; road conditions; and the locations of businesses, parking, and electric vehicle charging stations. Through the portal, partners are also able to request access to data tools and standards, and have access to the partners' forum, in which they can communicate directly with other partners and share ways of working with the data.[110]

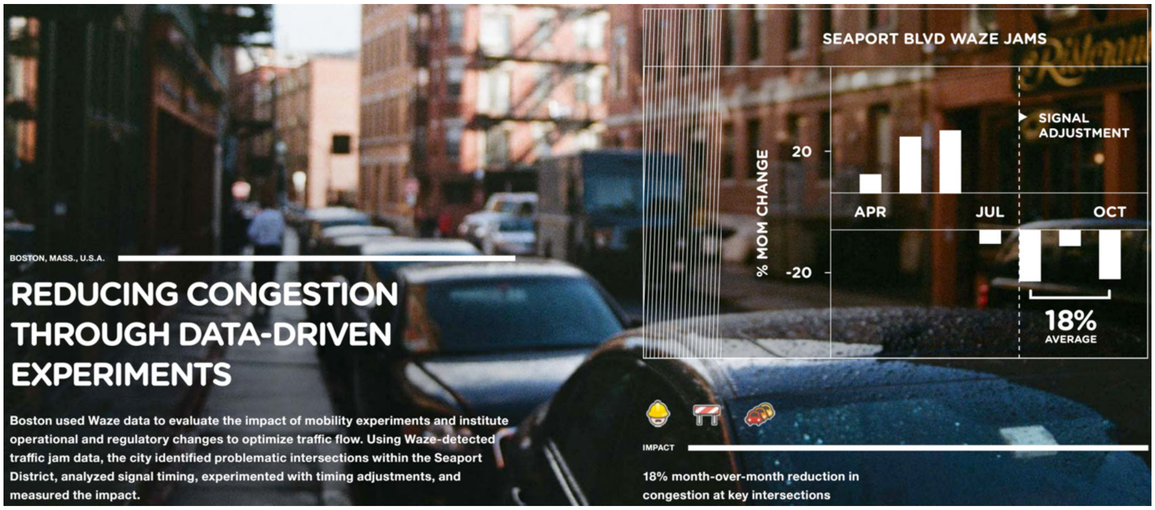
IMPACT

Most immediately, partners benefit from the CCP through faster incident reporting. According to one case study, Waze users reported crashes 4.5 minutes before the incident was reported through 911 or its local equivalent.[111] Waze's mapping capacity allows accurate locating of incidents, improving response and clearing times and potentially saving lives.[112] According to the FCC, reductions of one minute in ambulance response time can translate to over 10,000 lives saved annually in the US.[113] City partners like Boston have been able to integrate this incident reporting into existing traffic management systems to create a single real-time map view used in their traffic management center.[114]

CCP has several higher-level benefits to data collection and planning. The portal offers partners access to Waze's data standards for closure and incident reporting, to reduce data fragmentation and promote aggregation. It offers insight into hotspots with frequent congestion or incidents to promote better urban and transport planning. One of Waze's early partners, the city of Boston, was able to use CCP to identify problematic intersections in one district, analyze and experiment with signal timing, and measure its impact, to achieve an 18 percent reduction in congestion.[115] The Waze data enabled them to analyze not only the effect on the intersections themselves, but the follow-on effects to surrounding streets and intersections, to ensure the signal timing changes were not simply redistributing congestion. Boston also used the Waze data to assess the effects of transit signal priority for buses on congestion on streets crossing the bus routes, and the impact of construction on congestion and potential mitigation for this.[116]

Finally, Waze hosts partner summits and an online forum to encourage cross-pollination of ideas between partner cities and organizations,[117] a functionality that Boston CIO Jascha Franklin-Hodge describes as hugely valuable for moral support, idea sharing, inspiration, and learning from one another's mistakes.[118] Rio de Janeiro visited Boston to learn about its data-driven traffic planning and subsequently used Waze data before and during the 2016 Summer Olympics to guide building of new highways for Games-related traffic. During the Games, the city used a new API with Waze to automatically update the WAZE map with more than 430 road closures throughout the city, producing a 24 to 27 percent decrease in congestion during the morning commute.[119]

Although the data itself is beneficial, many city planners are deeply skeptical of the ability of apps like Waze to benefit cities. At its core, Waze is a navigation app for drivers, intended to minimize their commuting time by using all available roads, including residential side streets, to avoid traffic bottlenecks. In Fremont, CA, Waze users began to take to small streets in residential neighbourhoods to avoid freeway congestion. When chief transportation engineer Noe Veloso contacted Waze, he was told that the only solution was to change traffic rules to make side streets less attractive. He instituted commute-hour turn restrictions on the most heavily-used cut-through routes, but believes this may only have moved the problem to other neighborhoods.[120]



Boston used Waze data to evaluate the impact of mobility experiments and institute operational and regulatory changes to optimize traffic flow. Using Waze-detected traffic jam data, the city identified problematic intersections within the Seaport District, analyzed signal timing, experimented with timing adjustments, and measured the impact.

18% month-over-month reduction in congestion at key intersections

FIGURE 22

Franklin-Hodge acknowledges and shares these concerns about the effects of real-time navigation apps and the need for policy and regulatory measures to mitigate them, but stresses the importance of separating the data from the apps: "There's a set of data that Waze has opted to make available that we can and should use to help us better manage the road system. Non-engagement is an inappropriate and frankly unhelpful response to disruption you don't like."[121] Describing this kind of data use as "emerging best practice"—with Uber and Lyft similarly sharing data to inform urban planning[122]—he describes Waze's decision to make this data available free of charge to cities as progressive and praiseworthy:

> "Waze deserves a tremendous amount of credit for recognizing the potential partnership opportunities with cities and municipalities, and for actively exploring them. Most companies trying to disrupt or innovate in transportation are looking to keep government at arm's length or to extract additional revenue from government...[W]e are trying to evolve the transportation system, and therefore we are part of a greater whole that has a significant governmental component. If we want to be a positive participant in that system, we should be reaching out to people in government and helping them and finding ways that we can work with them to improve the quality of our service and the impact it has on communities. I'm optimistic that [this kind of approach] is actually the future as these companies go from early-stage disruptors to critical parts of the transportation ecosystem, and I hope that government is there to meet them."[123]

FACEBOOK INSIGHTS FOR IMPACT ZIKA

Featured Statistic or Number: In 2016, 58 percent of Facebook posts in Brazil related to Zika were created by men
Type of Data Collaborative: Intelligence Product
Sector: Health
Region: Brazil

PROBLEM STATEMENT

During a health crisis like the 2016 Zika epidemic in Brazil, health organizations and governments must understand the public conversation in order to create effective public information campaigns with the widest possible reach. In the past, the only way this information could be gained was through surveys and focus groups, which are expensive to conduct and often too slow to deliver insight. Since the expectations of the survey or focus group organizers shape the questions asked and thus the answers received, their results may also obscure important insights that could lead to a more effective campaign.

RESPONSE

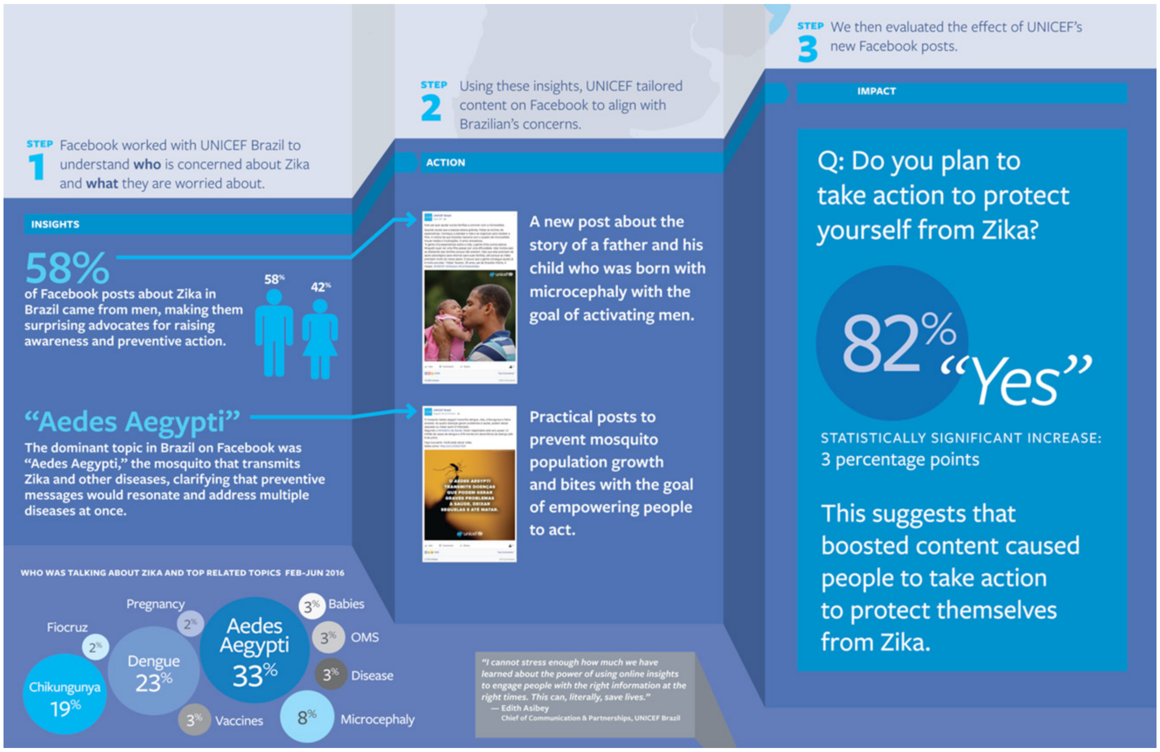
Social media gives inexpensive, near real-time access to public discourse on health crises through surveys and data analysis of user-generated content,[124] but also a vehicle for delivering timely, relevant and efficient public service announcements to a wide audience that may not access traditional media.

Aware that 90 percent of Brazilians use Facebook every month, Facebook and UNICEF Brazil collaborated to determine whether insights gleaned from data analysis of anonymized Zika-related Facebook posts could help UNICEF's social media awareness campaign to convince more Brazilians to take action to protect themselves from the virus.[125] Between May and August 2016, Facebook and data consultancy ActionSprout conducted surveys of Brazilian Facebook users and analyzed their user-generated Zika-related content to better understand the public conversation. The two organizations evaluated the data to find out who was posting and sharing information about Zika and what they were sharing, to determine what their concerns were. Their study then looked in detail at women aged 18 to 24, who would be most at risk of experiencing pregnancies affected by Zika. After analyzing the results, UNICEF Brazil launched a series of data-driven campaigns based on this information.[126]

OUTCOME/DELIVERABLE/PRODUCT

Facebook analyzed 17.3 million aggregated and anonymized interactions between May 5 and August 31, 2016, comprising 1.8 million posts, 13 million likes, 1.6 million comments and over 850,000 shares.[127]

At the outset, UNICEF and Facebook had assumed that women of childbearing age would be most concerned about Zika, and hence most active in posting and sharing about it. The data did indeed reveal that the most at-risk population for Zika by age was also the most engaged with Zika-related content on Facebook: 66 percent of posts were created by Brazilians between the ages of 18 and 34.[128] However, the data analysis also yielded a surprise: women aged 18 to 24 were the least likely to be posting and sharing about Zika—less likely than those over 35—while men aged 18 to 24 were most likely to post and share.[129] Overall, 58 percent of Facebook posts related to Zika were created by men.[130]



STEP 1 Facebook worked with UNICEF Brazil to understand **who** is concerned about Zika and **what** they are worried about.

INSIGHTS

58% of Facebook posts about Zika in Brazil came from men, making them surprising advocates for raising awareness and preventive action.

"Aedes Aegypti" The dominant topic in Brazil on Facebook was "Aedes Aegypti," the mosquito that transmits Zika and other diseases, clarifying that preventive messages would resonate and address multiple diseases at once.

WHO WAS TALKING ABOUT ZIKA AND TOP RELATED TOPICS FEB-JUN 2016

STEP 2 Using these insights, UNICEF tailored content on Facebook to align with Brazilian's concerns.

ACTION

A new post about the story of a father and his child who was born with microcephaly with the goal of activating men.

Practical posts to prevent mosquito population growth and bites with the goal of empowering people to act.

"I cannot stress enough how much we have learned about the power of using online insights to engage people with the right information at the right times. This can, literally, save lives."
— Edith Asibey
Chief of Communication & Partnerships, UNICEF Brazil

STEP 3 We then evaluated the effect of UNICEF's new Facebook posts.

IMPACT

Q: Do you plan to take action to protect yourself from Zika?

82% "Yes"

STATISTICALLY SIGNIFICANT INCREASE: 3 percentage points

This suggests that boosted content caused people to take action to protect themselves from Zika.

FIGURE 23

The analysis also examined what the top co-occurring terms were in Brazilian posts about Zika to understand the context in which Zika was being discussed. According to Drew Bernard, CEO of ActionSprout, the Zika study quickly revealed that Dengue and Chikungunya were the top co-occurring terms among Facebook posts about Zika, indicating a high level of public concern about mosquito-borne illness in general. Content quality and accuracy were also higher in posts that concerned all three mosquito-borne diseases than in those that discussed Zika alone. Subsequent focus groups confirmed this finding, allowing UNICEF Brazil to tailor its campaign to increase impact.[131]

These results point to one of the strengths of social media analysis as a tool for data-driven campaigning. Unlike a survey or focus group, social media analysis takes as its starting point the conversations people are already having. UNICEF Brazil began the campaign process expecting that women aged 18 to 24 would be their target audience in a campaign focused on Zika alone, but a campaign based on these assumptions would have failed to engage the audience in the way they hoped. Guided by social media insights, UNICEF Brazil was able to create content specifically geared to engage fathers in their social media information campaign, and to link Zika to other mosquito-borne illnesses to broaden their reach and generate greater audience trust.

IMPACT

Through its campaigns, UNICEF Brazil was able to more than double engagement, from 3.1 percent to 8.4 percent. In addition, people who saw these campaigns were more aware of UNICEF Brazil's efforts around Zika, with awareness rising from 29 percent to 36 percent. Most significantly, there was a three percentage point increase in those who said that they planned to take action to protect themselves from Zika, from 79 percent to 82 percent,[132] demonstrating that data-driven, high quality social media ads can save lives. This lift in intent was statistically significant at the five percent level, and outperformed 90 percent of studies conducted on Facebook ad performance.

PREDICTION AND FORECASTING

Social media data collaboratives can inform prediction and forecasting by enabling analyses into the data collected and held by online businesses. Through the use of data mining and machine learning techniques, researchers can build models to detect patterns in the thousands of data points produced each day, and create predictions and forecasts to improve the likelihood of success for new interventions and provide value to the public.

There have been many promising results across the board: in this section, we examine three initiatives in particular: two leveraging social media data in the health sector and the other focused on disaster response.

Case	Type of Data Collaborative	Sector	Region
Tracking the Flu Using Tweets	API; Prizes and Challenges	Health	United States
Predicting Floods with Flickr Metatags	API	Crisis Response	United Kingdom
Predicting Adverse Drug Events by Mining Health Social Media Streams and Forums	API	Health	United States, Spain

FIGURE 23

TRACKING THE FLU USING TWEETS

Featured Statistic or Number: 29 percent increase in accuracy when combining Twitter data with existing CDC flu data
Type of Data Collaborative: API; Prizes and Challenges
Sector: Health
Region: United States

PROBLEM STATEMENT

Though the flu mortality rate dropped to just 4,000 deaths in 2016,[133] it can still be a problem in the United States. American public health officials rely on regular reports released throughout flu season by the US Center for Disease Control and Prevention (CDC) in order to mitigate the virus's effects for a given year. Flu season varies greatly from year to year, and timely information on which strains of the virus are most active and to what degree allows officials to promote preventative measures most effectively.[134] Uncertain expectations for the upcoming flu season are responsible for losses in credibility for public health efforts, wasted and misdirected resources and, ultimately, increased sickness and mortality.[135] Due to natural delays inherent in the laboratory testing required, however, the most respected resource for influenza monitoring, namely the CDC's weekly **FluView** report, is always released with a one**-** to two-week delay.[136]

RESPONSE

The early 2010's witnessed a flurry of research on the use of social media in the analysis of public health issues, as epitomized by the widely cited 2011 paper by Johns Hopkins researchers "You are what you Tweet: Analyzing Twitter for public health."[137] Social media resources represented a unique opportunity to gauge public **behavior** and sentiment through real-time data. Also very attractive to researchers was the public nature of platforms such as Twitter, which contrasted sharply with earlier efforts by Google to tackle the same problem using their own private dataset of search queries and an internal team of researchers [See Box 1]. Researchers obtained their data for the most part through publically available web scraping methods.

The flu quickly became the most popular target of these efforts for a variety of reasons.[138] The large magnitude of flu cases every year resulted in more data available for analysis. In addition, the widely varying but annually occurring seasons of peak activity and the reliable CDC data available every week for verification served as compelling targets for predictive analytic models.

The first breakthroughs involved effectively screening out the useful tweets from the noisy ones—a process undertaken by researchers including Professor Dredze of Johns Hopkins.[139] A tweet lamenting a user's own affliction with the virus is far more useful than another tweet hoping Lebron James's case of the flu will not keep him off the basketball court for long. The Johns Hopkins team was also the first to explicitly measure the predictive power of the model they devised.

OUTCOME/DELIVERABLE/PRODUCT

These efforts by researchers resulted in a plethora of predictive models relying in some part on social media data to predict upcoming activity during the flu season. The model developed by the Johns Hopkins researchers was capable of nowcasting flu conditions with a 29 percent increase in accuracy by combining Twitter data with the previously available CDC data.[140]

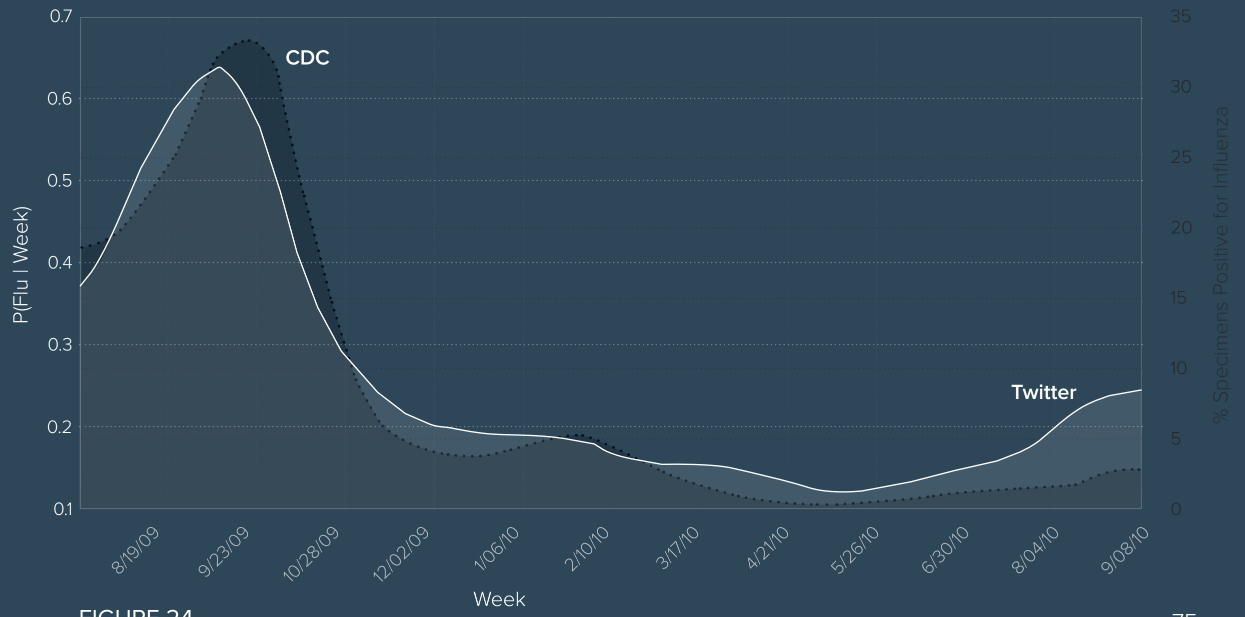


FIGURE 24

In 2013, the CDC, having taken notice of this increased interest in social media flu prediction, organized and hosted its first "Predict the Flu Season" challenge.[141, 142, 143] This challenge focused the nine participating teams of researchers on predicting various milestones of the influenza season—e.g., start, duration and peak—which would be the most useful to public health prevention efforts.

IMPACT

The model developed by the researchers at Johns Hopkins clearly demonstrates the promise and potential of the effective use of social media data to inform flu prediction. Their influenza model could predict the virus's future activity one to two weeks before the CDC data alone with the same accuracy. There has been relatively little evidence that these breakthroughs are being widely used in an official capacity, however. Moreover, a 2016 paper by the CDC summarizing the results of its first flu prediction challenge stated that the science was still in need of improvement before it could achieve a transformative impact.[144]

THE SHORTCOMINGS OF GOOGLE FLU TRENDS

The promise of using data gleaned from the Internet to predict the flu first gained widespread attention with Google Flu Trends (GFT), an effort by the Internet search company to analyze user search queries to forecast the number of influenza cases throughout the year. Although at first praised for its surprising similarity to officially released CDC data, the service would become the subject of criticism for its failure to foresee the intensity of the 2009/2010 H1N1 flu pandemic that rocked the United States. The Google researchers reportedly made adjustments to their algorithms following this failure. Criticism only increased however, after a 2015 article published in ***Nature*** brought attention to GFT's consistent overestimation of flu cases within the previous two years. At one point GFT's had been predicting twice as many doctor visits due to influenza than had actually occurred. GFT was officially shut down in 2015 although past data is still publically available for download. Researchers noted throughout GFT's lifespan that the model's results were utterly unreproducible thanks to a lack of transparency on the data and algorithms involved behind the scenes.[145]

PREDICTING FLOODS WITH FLICKR METATAGS

Featured Statistic or Number: The YFCC100M dataset contains 14GB of information and metadata from the platform
Type of Data Collaborative: API
Sector: Crisis Response
Region: United Kingdom

PROBLEM STATEMENT

In a flood situation, every moment counts, and delays in the mobilization of early warning systems or relief efforts could be the difference between life and death. After witnessing the disastrous effects of flooding events around the United Kingdom resulting from 2013's Cyclone Dirk and Storm St. Jude, researcher Nataliya Tkachenko was inspired to explore ways data could improve the effectiveness of early warning systems and predictive efforts surrounding flood events more generally.[146]

RESPONSE

To improve prediction around flooding events, **Tkachenko** and her collaborators sought to "reconstruct the entire social landscape, in all its dynamics, from the very beginning of the event toward its very end." More specifically, they developed an analytical methodology using social media data to identify and cluster keywords (e.g., 'river,' 'flood,' 'overflow') related to emerging events, and used the technique to gain a better understanding of flood conditions. Beyond just monitoring events in real-time, a key focus for the project was improving the ability to use "alternative environmental semantics"**—**i.e., social media data—to "serve as a predictor of an evolving hazard."[147]

The project is part of a broader move toward using user generated content (UGC) to help develop warning systems. The U.S. Geological Survey, for example, sometimes uses Twitter data as a type of social sensor.[148] Similarly the FloodTags platform shares a dashboard, API, and software development toolkit with diverse actors to monitor Twitter postings related water problems.[149]

Tkachenko also began her experimentation with social media for flood prediction with the use of Twitter data. Early experimentation with data from the platform, however, made clear that finding the signal in the noise regarding flood events required a high level of data cleaning—e.g., "filtering out re-tweets, tweets outside flooded areas, even tweets about different flood events mentioned at the time of the current one."[150]

After experimenting with Twitter data, Tkachenko came to realize that images provide a strong verification mechanism ensuring relevancy, and, as a result, she turned to Yahoo!'s Flickr platform, which is built on user-uploaded photographs and videos. The flood prediction approach was built on data from the Yahoo! Flickr Creative Commons 100M (YFCC100M). The YFCC100M dataset contains metadata from photos and videos uploaded to the platform from April 2004 to August 2014. This same Flickr dataset was previously used to study air pressure trendlines during the period when Superstorm Sandy affected many areas across the Atlantic.[151]

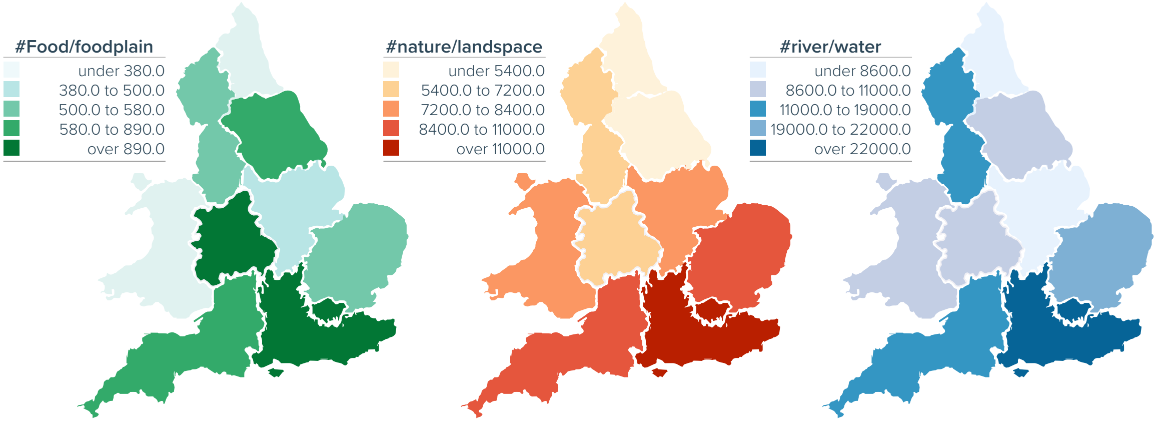


FIGURE 25

The researchers' analytics focused on the Flickr images themselves, the captions included, and related metadata, including timestamps and geolocation data. Given the fact that many photos taken in a flooding context are snapped using smartphones, the geolocation and timing metadata can enable 'nowcasting' or 'present-time reporting' since, given certain app settings, Flickr's uploading and tagging of images is nearly simultaneous to shooting the photo.

OUTCOME/DELIVERABLE/PRODUCT

In the short term, the outcome of the research was a published research paper. The paper articulates a new approach for leveraging social media in a flooding situation that goes beyond previous efforts using Twitter and other datasets. In the longer term, the research is intended to act as a blueprint for a social media data-driven "nowcasting tool" and early response system to improve awareness of current conditions and enable prediction. Eventually, the intended outcome is leveraging a diverse range of data, including social media and sensor data, to "detect an event before its outbreak."[152]

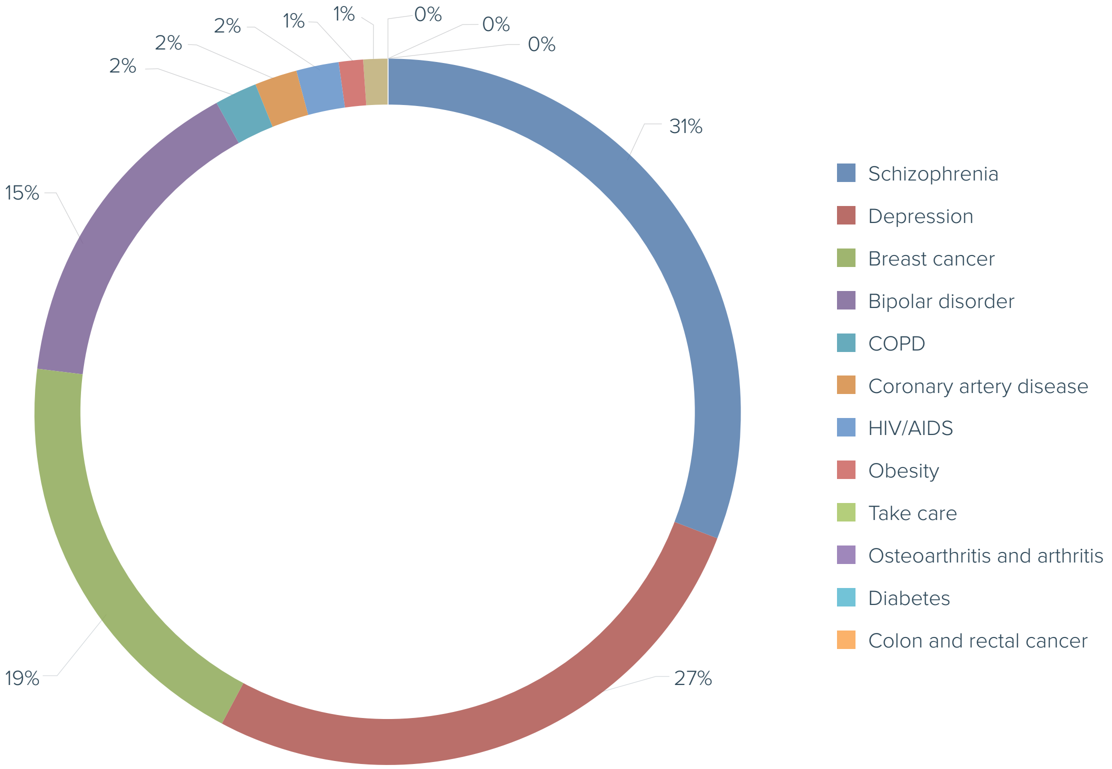


FIGURE 26 – DISTRIBUTION OF USER COMMENTS

IMPACT

While the method articulated and tested in this research is still being honed and developed, the initiative has created an increase in knowledge and is a self-proclaimed "substantial leap" forward in the application of social media data analytics to better understand the natural sciences. The initiative showed that analyzing how flooding-related keywords on social media "temporally relate" can aid in the identification of early stages of a flood event before traditional systems of identification and prediction are aware of the threat.[153]

In particular, the research shows how social media data analysis could work in concert with hydrological and meteorological models to create improved early warning systems and enable improved resilience planning. The initiative also demonstrated how a learning algorithm applied on the existing corpus of data could be used in a similar way in real-time, aggregating and interpreting real-time data from social media, sensors and other sources to help raise awareness of current conditions and improve forecasting capabilities.

Looking ahead, while the data provided in the YFCC100M dataset enabled useful analysis, a more collaborative arrangement between Yahoo! Flickr could improve the utility of such efforts. Tkachenko notes that, while "Yahoo! did a tremendous job on making such as huge archive (100M) available in the public domain," the public value of her research would improve with access to the "complete albeit anonymized data entries in order to get spatio-temporal risk mapping as complete as possible."[154] Data with this level of personally identifiable information would create clear privacy risks if made fully accessible like the YFCC100M, but more targeted sharing arrangements with identified partners could amplify impacts.

Next Tkachenko hopes to take the lessons learned in this effort and experiment with larger scale analytics work seeking to construct critical clusters of risk-related words to help inform strategic work across political, social, cultural, and other emerging events.[155]

PREDICTING ADVERSE DRUG REACTIONS BY MINING HEALTH SOCIAL MEDIA STREAMS AND FORUMS

Featured Statistic or Number: Social media data analytics would have identified certain adverse drug reactions up to 9 years before the FDA identified them
Type of Data Collaborative: API
Sector: Health
Region: United States, Spain

PROBLEM STATEMENT

Adverse drug reactions (ADRs)—"unwanted, uncomfortable, or dangerous effects that a drug may have"[156]—are a leading cause of death among hospitalized patients. As it stands, clinical trials are the primary mechanism through which ADRs are identified, but clinical trials arelimited in terms of scope and potential participants included. For instance, a clinical trial could neglect to include patients with particular chronic illnesses that could interact with medications in a way that introduces ADRs.[157] As a result, ADRs often slip through the cracks during clinical trials, and are not reported until, in the U.S. context, the Food and Drug Administration (FDA) issues an official drug label change.[158] To help the FDA, or the European Medicines Agency (EMA) identify ADRs not recognized in clinical trials, healthcare professionals are asked to voluntarily report ADRs as they occur. This addition to the traditional healthcare provision workflow and a reluctance to report ADRs unless there is absolute certainty often leads to under-reporting of ADRs. Official patient reporting systems for ADRs also exist, but are not used to a significant degree.[159]

RESPONSE

To help improve the ability of government drug regulators to identify ADRs after the clinical trial period and communicate more effectively with patients regarding potential risks and side effects, a number of researchers are looking to social media platforms for emergent evidence of ADRs.

A project in the U.S., for example, created an extraction and processing methodology that parsed 100,000 messages from popular medical forums to evaluate the likelihood of an ADR to occur beyond simple chance.[160]

Another project built a natural language processing and machine-learning approach to identify ADRs through analysis of user reviews of drugs in health social networks like DailyStrength. Focusing only on four drugs—carbamazepine, olanzapine, trazodone, ziprasidone—the researchers leveraged a corpus of 6,890 comment records to experiment with these data-driven techniques to identify ADRs. The four drugs were chosen due to the existing knowledge regarding their ADRs, allowing the accuracy of the NLP approach to be verified against existing ADR research and analysis.[161]

Researchers at Carlos III University of Madrid leveraged data from the Spanish-language social network Forumclinic to determine if "health-related social media can be used as a complementary data source to spontaneous reporting systems in order to detect unknown ADRs and thereby to increase drug safety."[162] The project began by focusing on Twitter, but they quickly learned that conversations around ADRs rarely occurred on the platform. They similarly explored using Facebook but found that relevant conversations tended to happen in closed groups where information could not be scraped. Forumclinic, on the other hand, was an open platform, with high levels of use among relevant communities.[163]

OUTCOME/DELIVERABLE/PRODUCT

The two U.S.-focused efforts' primary deliverables were published research papers demonstrating the feasibility of such a social-media driven approach to ADR-detection.

Source: ***Detecting drugs and adverse events from Spanish health social media streams***

For the Spanish effort, in addition to published research, a collaborative annotation effort yielded a corpus of data comprising 187 drugs and 636 adverse events. This reusable 'dictionary' of how drugs and ADRs are referenced in online discussions was then opened for academic purposes and future quantitative research.

IMPACT

The research efforts exist primarily as proof-of-concept for the use of social media data to improve the detection of ADRs after drugs move past their clinical trial period. The methodology used in the Forumclinic initiative, for example, was 87 percent accurate in identifying drugs in posts on the platform, and 85 percent accurate in identifying ADRs.[164] Two specific verification exercises further demonstrate the potential impact of such a predictive approach. The first case study showed that using social media data in a predictive capacity would have identified ADRs related to cognitive impairment a full nine years before the FDA added the issue to the labels for medications in the Statins drug class. A similar exercise using a sample made up of social media data from 1999 to 2008 found that the use of the researchers' methodology would have captured the prevalence of agitation ADRs among users of Wellbutrin in 2002; the FDA included agitation on Wellbutrin's label for the first time in 2009.[165]

In addition to mobilizing this type of approach, researchers are looking to expand and automate the types of analyses that can be performed. This effort is targeted, in particular, at "creating structured summaries (for human consumption) of all adverse effects experienced and reported by patients during the lifetime of a drug."[166]

IMPACT ASSESSMENT AND EVALUATION

Finally, social media data collaboratives can be used to assess and evaluate the impact of various policies or interventions. Through a careful analysis of social media activity using target words and phrases as filters, researchers and policymakers can gain deeper insights into public opinion and behavior as they relate to particular initiatives. The final three case studies look into the swirl of public discussion around campaigns and policies across the globe: the first seeks to understand the impact of a campaign to break down gender barriers, while the last two monitor social media engagement surrounding policy changes.

Case	Type of Data Collaborative	Sector	Region
Sport England's This Girl Can	API	Health	United Kingdom
Using Twitter Data to Analyze Public Sentiment on Fuel Subsidy Policy Reform in El Salvador	API, Research Partnership	Energy	El Salvador
Measuring Global Engagement on Climate Change	API	Climate	Global

FIGURE 27

#THISGIRLCAN: BREAKING DOWN THE GENDER GAP IN ENGLISH SPORT

Featured Statistic or Number: 1.6 million women started exercising as a result of the campaign
Type of Data Collaborative: API
Sector: Health
Region: United Kingdom

PROBLEM STATEMENT

Sport England, a government-funded agency with a remit to engage citizens in sport for life by creating opportunities for participation nationally and locally, identified a persistent gender gap in sport participation between men and women.[167] This shortfall, equating to 2 million fewer women aged 14 to 40 than men taking part in sport regularly,[168] persisted despite the effects of the health and fitness industry and the spike in sporting participation produced by the London 2012 Olympic Games.[169]

RESPONSE

Sport England developed a strategy to launch what they have described as their "sassiest campaign ever" to address the gender gap in sport in England. In 2014, in addition to talking to stakeholders from major league football clubs to new mothers, Sport England **analyzed** over 10 million posts by women on sport and exercise on social media during 2014.[170]

Social media analysis using Crimson Hexagon's social listening tool, ForSight, revealed that while women looked forward to and enjoyed exercise and sport, and felt strongly positive afterwards, three barriers to exercise affected different groups of women. Young women (14 to 24 years) were put off by their appearance or how they looked while exercising. Mothers felt guilty both when exercising because they were not spending time with their families, and when they did **not** because they felt they were not being good role models. Other women either felt that they were not good enough to participate, or were too good and would make others uncomfortable.[171] According to Jennie Price, chief executive of Sport England: “One of the strongest themes was a fear of judgment.”[172] Social media analysis allowed the campaign to surface the emotional barriers that kept women from exercising.[173]

OUTCOME/DELIVERABLE/PRODUCT

Sport England set out to develop a campaign with a positive, encouraging and empathetic tone that connected with women's concerns, built around the manifesto: “Women come in all shapes and sizes and all levels of ability. It doesn't matter if you're a bit rubbish or an expert. The point is you're a woman and you're doing something.”[174] According to Kate Dale, Sport England's Head of Brand and Digital Strategy, all of the campaign's messaging came out of social media analysis and talking to women. “Our key point is staying where the women are, meeting them where they are.”[175]

FIGURE 28

Sport England began reaching out to influencers and building a social network of women already actively talking about sport and exercise on Facebook and Twitter.[176] Having built this online community, the campaign went live in 2015 with several forms of advertising, including a 90-second TV ad spot in a major soap (Coronation Street) which was viewed by over 7.5 million people on TV,[177] and more than 37 million times since on Facebook and YouTube.[178] The campaign's ads have all featured real women "streetcast" from gyms, parks, football pitches, and pools, with encouraging messaging such as: "I'm slow but I'm lapping everyone on the couch."[179] The campaign is ongoing, and a subsequent ad in early 2017 has targeted women over 40 with a voiceover of Maya Angelou reading "Phenomenal Woman."

An established community of supporters ready to amplify the campaign message allowed Sport England to make the most of a media buy limited by public funding: "We knew that we needed a community that took this to heart and would spread the message for us if we were going to be able to achieve what we wanted to achieve, but also to ensure that we got value for our money. And it's social [media] that's been the beating heart of the campaign. If we'd just had the TV ad we wouldn't have had the same results."[180]

According to Kate Dale, social media has also allowed the campaign to support women in day to day decisionmaking about exercise. "You're able to get close to women at the point at which they're having to decide, am I going to the gym today? You need to be there at that moment and social allows you to do that in a way that TV doesn't."[181] The campaign's active social media community of 500,000 has a tailored algorithm that sends encouraging tweets to women tweeting about exercise or sport. In addition to Facebook and Twitter, the campaign has begun to use other outlets including Medium for longer-form content, Instagram, and Snapchat, chiefly to reach younger women and teens.[182] The campaign has made subsequent forays into commercial partnerships for merchandise, and a side campaign which encouraged women to raise money for the charity Sport Relief while exercising.

IMPACT

As a result of the campaign, 1.6 million women started exercising, and the number of women playing sport and exercising is increasing faster than that of men. Furthermore, the gender gap in participation has begun to narrow (from 1.78 million to 1.73 million).[183] The Media tracking revealed that over one in three women who had viewed the campaign claimed to have increased their activity levels as a result.

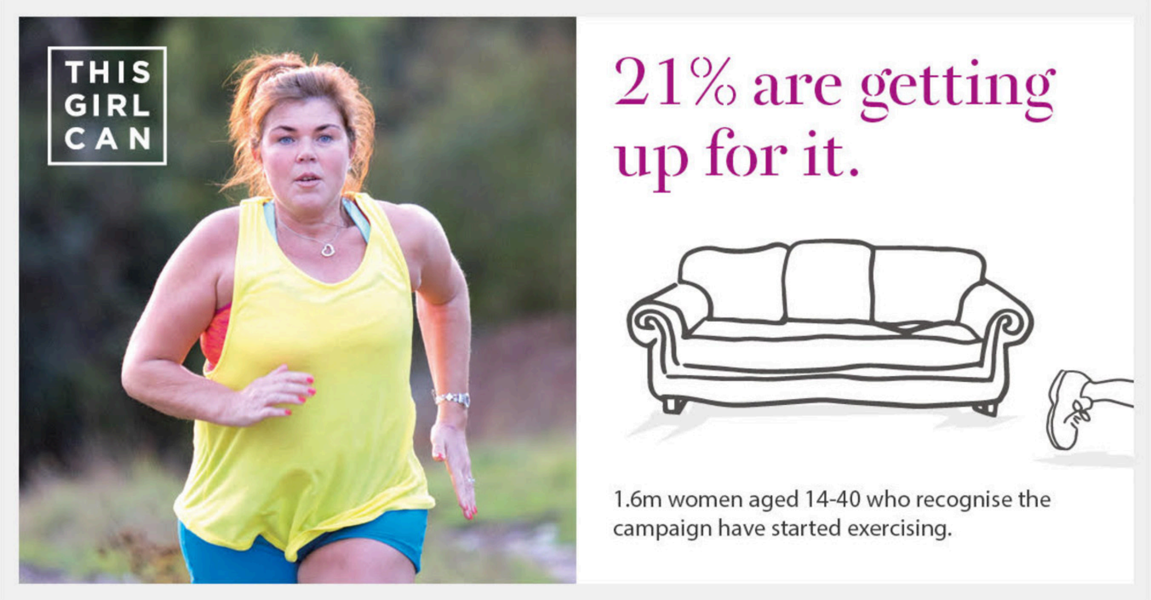

THIS GIRL CAN
21% are getting up for it.
1.6m women aged 14-40 who recognise the campaign have started exercising.

FIGURE 29

As a result of social media analysis of the campaign's impact, Sport England realized that they would need to change how they viewed the campaign's purpose.[184] "In the early days we talked about liberating women from the fear of judgment. That's a lovely aim, but I think it was a bit grandiose."[185] In addition to measuring participation, Sport England continues to measure women's confidence and sense of belonging while exercising. As Kate Dale notes, both are increasing, but much more slowly than participation rates.[186] While the campaign was a resounding success, achieving 100 percent exposure in its target audience and prompting 2.8 million women to increase their activity as a result of it, it has not dramatically changed attitudes: "That fear of judgment is still there. What we are doing is helping women manage that fear and do it anyway."[187]

USING TWITTER DATA TO ANALYZE PUBLIC SENTIMENT ON FUEL SUBSIDY POLICY REFORM IN EL SALVADOR

Featured Statistic or Number: By September 2013, 10.1% of fuel subsidy-related tweets demonstrated a distrust in institutions, down from 24.2% in January 2011.
Type of Data Collaborative: API; Research Partnership
Sector: Energy
Region: El Salvador

PROBLEM STATEMENT

In April 2011, the government of El Salvador removed a universal subsidy on propane gas, the country's most commonly used domestic cooking fuel, replacing it with an income transfer to eligible households. The policy change more than doubled the cost of a 25-pound bottle of LPG from US$5.10 to US$13.60. The policy was highly unpopular, with a national survey in January 2011 showing only 30 percent of the population in favour of the reform, and sparked a series of strikes by companies that distributed the gas. This low approval rating increased slowly over the following 18 months before stabilizing at 65 percent in favor in September 2013.[188]

RESPONSE

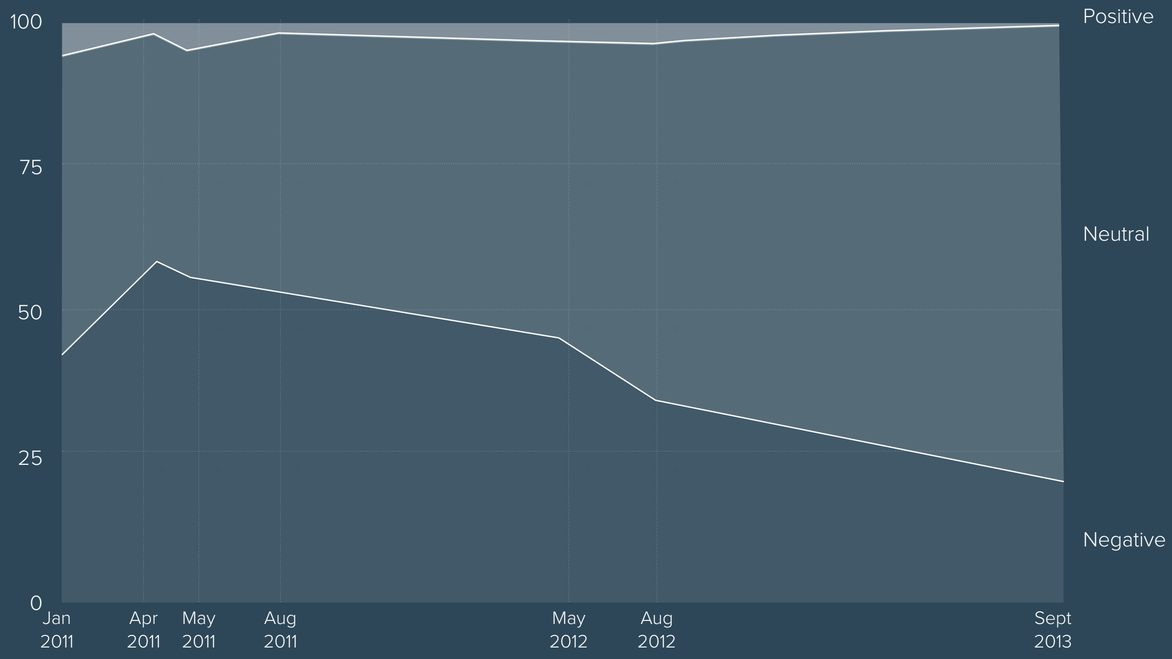
In the light of preliminary research on the role of partisanship, community information and distrust in institutions, Global Pulse and the World Bank collaborated to investigate whether social media data could be used to provide insights into public opinion as a complement to, or substitute for, household surveys.[189]

OUTCOME/DELIVERABLE/PRODUCT

With help from regional experts, a taxonomy of keywords, including local dialects and slang, was developed to filter Twitter for content related to the policy reform based on five pillars of public concern: lack of information, partisanship, distrust of institutions, personal economic impact and other, unclassified concerns. Tweets were filtered and a subset examined to confirm their relevance and further refine the taxonomy; this process was then repeated twice more until the team was satisfied that they had removed all but a very few irrelevant tweets. At that point, tweets were filtered again to restrict the sample to content from El Salvador. The project categorized tweets by content and sentiment (positive, neutral, or negative).[190]

IMPACT

The study confirmed the potential of social media data as a complement to household surveys, since the public sentiment observed in the Twitter analysis broadly echoed public opinion from the newspaper surveys. The Twitter analysis, however, was somewhat more nuanced: instead of showing increased satisfaction with the policy reform between January 2011 and September 2013, as the household survey had done, it showed a decline in negative sentiment and rise in neutral sentiment, but no increase in positive sentiment. Miguel Luengo-Oroz suggests that the capacity to reveal public sentiments such as anger, happiness, and fear, instead of a binary positive and negative, is one of the strengths of this technique. It also allows researchers to elucidate sometimes surprising relationships between topics by analyzing the word cloud surrounding a given topic, and to feed these insights almost immediately into campaigns and counter-campaigns.[191] It also showed the potential for social media as a near real-time measure of public opinion of policy reforms, especially where other measures are unavailable.[192]



The figure above shows the evolution in sentiment form January 2011 when the subsidy was passed to July 2013. Positive and negative sentiment gradually decreased, while neutral sintiment increased over the time period.

FIGURE 30

Since the completion of the project, UN Global Pulse has gone on to use the same methodology in other studies, including perceptions of HIV in Brazil.[193] Despite its usefulness, Miguel Luengo-Oroz cautions that social media cannot replace statistics and that it is inherently difficult, if not impossible, to control the demographics of the sample. He suggests that studies of this type are best used to fill data gaps between surveys and to act as an early warning system of emerging problems requiring deeper investigation.[194]

MEASURING GLOBAL ENGAGEMENT ON CLIMATE CHANGE

Featured Statistic or Number: 400,000 English-language tweets analyzed on the days of the People's Climate March and Climate Summit
Type of Data Collaborative: API
Sector: Climate
Region: Global

PROBLEM STATEMENT

In September 2014, the UN Secretary-General convened over 100 heads of state and world leaders in business and civil society for the first Climate Summit. Leaders made bold announcements on climate action, paving the way for a global climate conference to be held in Paris in December 2015. UN Global Pulse and the UN Secretary General sought a way to assess how much people globally were talking about climate change, and in what contexts, in order to identify citizen concerns and target communications efforts. They also wanted to establish a baseline level of engagement so that they could assess the impact of the Paris meeting and other climate-related events on public engagement, and examine how climate change was discussed regionally.

RESPONSE

UN Global Pulse and the UN Secretary General's Climate Change Support Team collaborated to develop a tool to monitor social media engagement in the months leading up to and following the Paris Climate Summit of 2015. They identified commonly used words and phrases to filter relevant tweets and assess spikes in conversation levels and their relationship to events.[195]

OUTCOME/DELIVERABLE/PRODUCT

The publicly accessible, online social media monitor[196] analyzed the volume and content of tweets about climate change in three languages (English, Spanish and French) each day. Using a taxonomy of some 1,000 words and phrases related to climate change, they filtered over 15 million tweets from April 2014. The tool also assessed tweets in eight related subfields—energy, economy, risk/disaster, agriculture/forestry, weather, arctic, and oceans/water—to extract relevant content. In addition to measuring volumes and topics of tweets, the tool measured hashtags, top accounts and domains used and referenced, top retweets, sentiment, country-level volume and content, and Google search trends (also by region).[197]

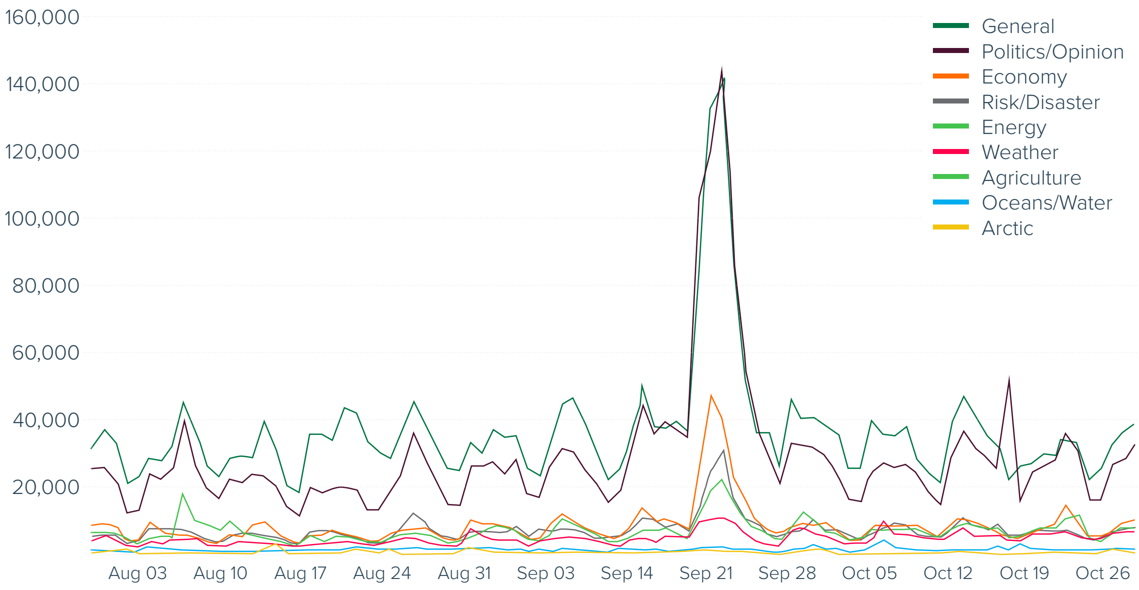


FIGURE 31 – THE GRAPH ABOVE SHOWS THE DAILY VOLUMES OF ENGLISH-LANGUAGE TWEETS ABOUT CLIMATE CHANGE AND VARIOUS RELATED SUB-TOPICS.

IMPACT

The social media monitor succeeded in providing a real-time measure of public engagement with climate change. Baseline volume was about 140,000 English-language tweets per day from the tool's inception in April 2014. On the days of the People's Climate March and Climate Summit, the volume of tweets rose to over 400,000. Three- to four-fold increases were also observed in Spanish- and French-language tweets. In the month following the Summit, the baseline volume rose by 10 to 15 percent, and positive public sentiment spiked on the day of the Climate Summit.[198]

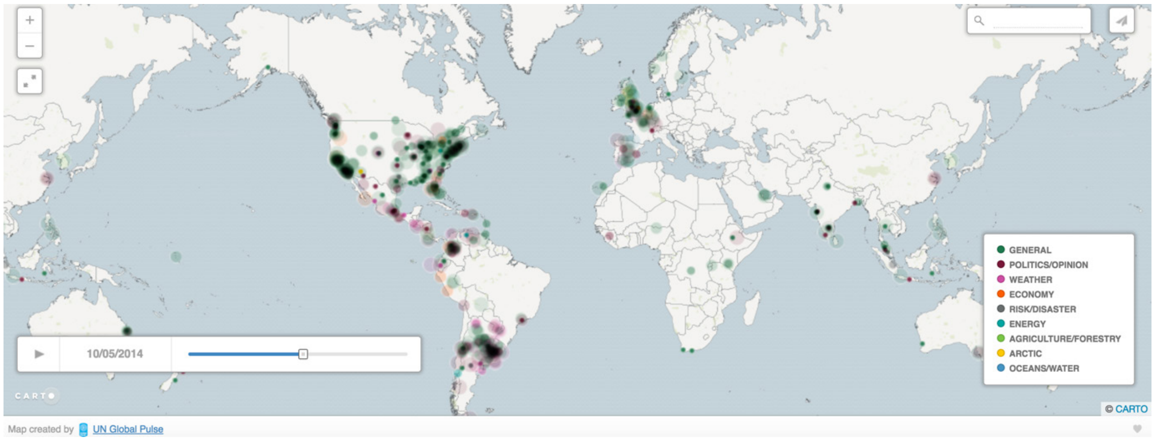


FIGURE 32

The monitor helped the UN to establish top influencers, who were often politicians, entertainers and business leaders, and to identify five media outlets whose Twitter accounts were most often mentioned and hence could be most effective in reaching engaged audiences. Since the monitor showed that Greenpeace was one of the most influential Twitter accounts before the Summit, the UN Secretary General's Climate Change Support Team engaged Greenpeace to post tweets supporting the Summit.[199]

Finally, the monitor exposed regional concerns and priorities which were then visualised in an interactive map displaying volume and issues. Such identification of regional concerns could be enormously helpful in advancing climate policy locally and developing locally relevant communications.[200] Miguel Luongo-Oroz also suggests that this kind of social media use has value as a means of "amplifying the voices of people" and indicating that administrations and organizations are listening to their constituents. At the same, he believes it's best at identifying and tracking short-term trends, so its value for long-term monitoring may be limited.[201]

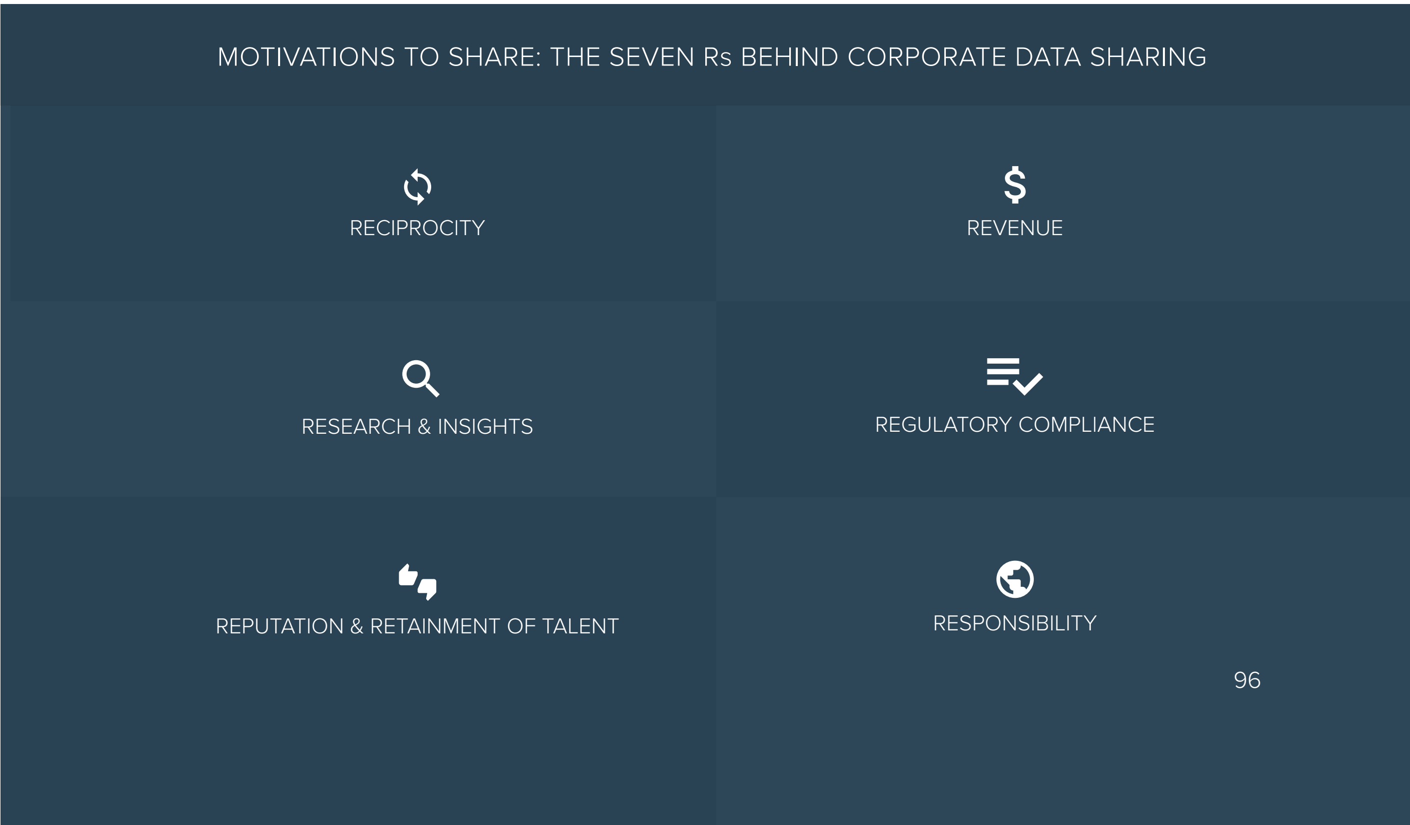
IV. RISKS AND REWARDS

Responsibly Sharing Social Media Data

MOTIVATIONS AND INCENTIVES FOR SOCIAL MEDIA CORPORATIONS TO SHARE DATA THROUGH DATA COLLABORATIVES

The preceding discussion has focused primarily on the public interest benefits of social media data collaboratives—i.e., what could be called the "data philanthropic" case. But there are also powerful business incentives for data collaboration that go beyond pure philanthropic motivations. In fact, the most successful data collaboratives are built around win-win propositions in which both the public and the private sector entities benefit. It is important to make this business case—and to address potential corporate and societal concerns over data sharing—in order to place data collaboration at the very center of social media companies' workflow. Only then can we ensure that data collaborative initiatives are not simply relegated to the status of one-off projects that can easily be sunsetted or ignored.

FIGURE 33



MOTIVATIONS TO SHARE: THE SEVEN Rs BEHIND CORPORATE DATA SHARING

RECIPROCITY

REVENUE

RESEARCH & INSIGHTS

REGULATORY COMPLIANCE

REPUTATION & RETAINMENT OF TALENT

RESPONSIBILITY

Based on our research, we have identified seven benefits for social media companies that can follow from data collaboration. We call these the Seven Rationales or 7Rs; they represent key incentives for the private sector to share social media data through data collaboratives to the end of creating social media intelligence:

- **Reciprocity:** Social Media Data Collaboratives can be designed in a way that enables reciprocity, where corporations share their data and also gain access to other data sources or domain expertise that may be important to their own business decisions.
- **Research Insights:** Analysis enabled by data sharing can uncover new research insights that might be useful to the organization sharing the data.
- **Reputation:** Sharing data for public good may enhance a firm's reputation and public image.
- **Revenue:** Data collaboratives can allow private sector entities to generate revenue by making their data available for a price. Sharing data doesn't always have to be free.
- **Regulatory Compliance:** In some regions and situations, data sharing can be required for regulatory compliance.
- **Responsibility:** Companies can derive value from social responsibility not just because of the positive image such an activity produces, but because opening up data can also improve the competitive business environment within which the business operates.
- **Retainment of Talent:** Working on problems that matter enables companies to recruit or retain data science talent.

DATA SHARING RISKS AND BARRIERS

It is not enough simply to show that data collaboratives offer benefits to companies. As noted, even companies that may want to share can encounter challenges in the form of internal hesitations, concerns or obstacles. Here, we address the four most frequently cited challenges, and attempt to explain how data collaboratives can help mitigate the embedded concerns or obstacles.

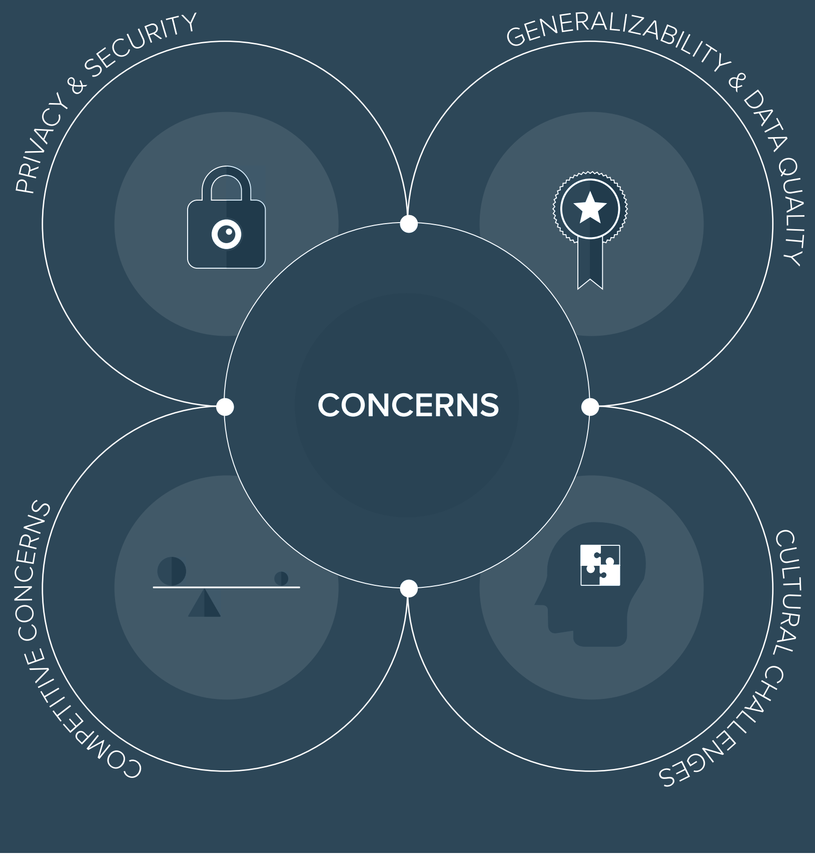


FIGURE 34

GENERALIZABILITY, DATA BIAS, AND QUALITY

A key concern when using social media data involves the level of representativeness[202] or data bias. Social Media Data often represents a particular demographic subset—possibly ignoring the so-called "data invisibles". Caution needs to be exercised in extrapolating general observations, with consequences for the population at large, from such data. Social media data is also inherently noisy, comprising information unrelated to the question at hand; and often dirty, containing factual errors or duplicative data. This results in the need for extensive data filtering and cleansing prior to extracting insights from the data. Additionally, the opaqueness of algorithms used to analyze data flowing through data collaboratives leads to an inability to reproduce or meaningfully scrutinize these analyses. As a result, biases and quality issues are more likely to go unnoticed.

PRIVACY AND SECURITY

The most common concern expressed by individuals and companies involves the worry that sharing information may result in disclosing personally or demographically identifiable information, affecting (group) privacy or creating security violations. Such concerns are not only natural, but very important: data sharing must not in any way result in dilution of protections for individuals, many of whom might not even be aware that the data was collected about them in the first place. Moreover, without adequate foresight and planning, existing privacy protections could be eroded as a result of the unforeseen availability of new datasets and/or data analysis techniques arising from data collaboratives.

COMPETITIVE CONCERNS

Data and the insights it offers are often at the core of social media companies' businesses. The prospect of sharing data—usually without charge—may therefore seem to threaten vital commercial interests or affect their competitive advantage. While certainly valid and important to consider when developing data collaboratives, our exploration above also indicates that this view begins from a false zero-sum-game understanding of data collaboration and its potential—ignoring the incentives and benefits of establishing data collaboration listed above.

CULTURAL CHALLENGES OF DATA COLLABORATION

Finally, our exploration indicates that one of the chief obstacles to more widespread data sharing—and one that may underlie the above discussed concerns—is simply a lack of awareness, comfort or practice with the notion of sharing social media data for good. Like much of the economy, the data economy exists within an ecology of well-defined (and, in our view, largely outdated) concepts like ownership, private property, and data hoarding. Yet social media companies today operate in a rapidly changing environment where notions of collaboration, sharing and mutual benefit are far more widely acceptable—both as philanthropic and commercial propositions. The MasterCard Center for Inclusive Growth, for example, provides a potential blueprint for formalizing data-sharing initiatives within corporations to help move away from one-off and easily abandoned projects and toward a codified institutional culture of sharing.[203] Placing such values at the heart of business operations is arguably one of the central challenges facing many social media companies today.

To be clear, the problem-driven data collaborative culture described in this paper should not be viewed as a replacement or upgrade on the current open data and open source movements. The continued move toward cross-sector data collaboration should not distract from important efforts to maintain access to publicly useful information, including open data.

TOWARD DATA RESPONSIBILITY FRAMEWORKS FOR SOCIAL MEDIA DATA COLLABORATIVES

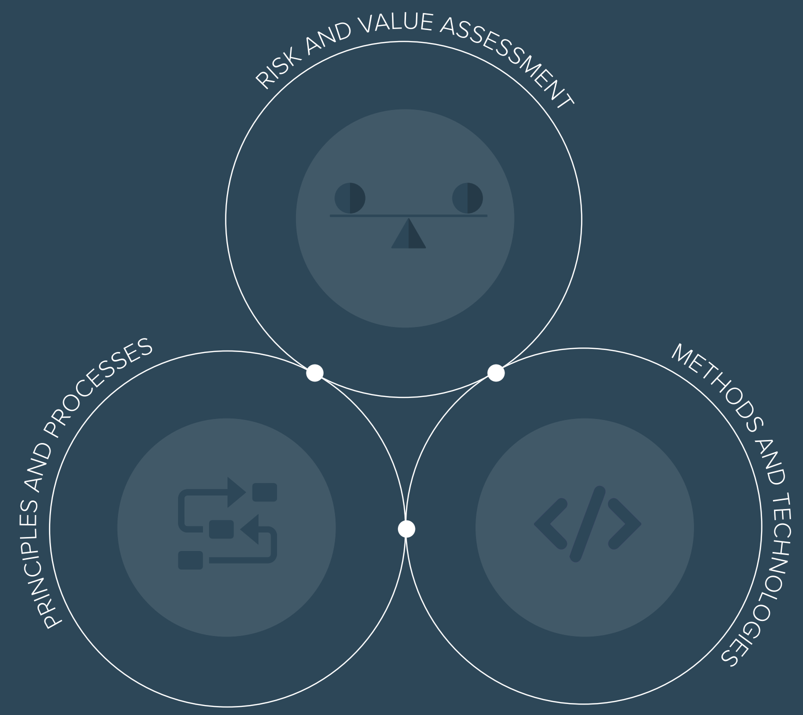


FIGURE 35

A number of steps can be taken to foster a culture of data collaboration within companies while addressing the concerns listed above. Several social media corporations and organizations—such as Facebook, the International Committee of the Red Cross and the United Nations Global Pulse, to name just a few—have started to explore various frameworks to share data responsibly, yet it is unlikely (and undesirable) that one single data responsibility framework will be able to handle all the context specific needs and risks associated with social media data sharing. At the same time, data responsibility frameworks put in place by public and private sector entities will likely need to include the following four elements to assure concerns are addressed meaningfully and legitimately. They include:

VALUE AND RISK ASSESSMENTS

Risks—including privacy, ethical and commercial concerns—exist across the social media data lifecycle, and range from inaccurate, non-representative data entry at the collection stage; to insufficient, outdated, or inflexible security provisions at the processing stage; to incompatible cultural or institutional norms or expectations at the sharing stage; to aggregation or correlation of incomparable datasets at the analysis stage; to controversial or incongruous data usage.

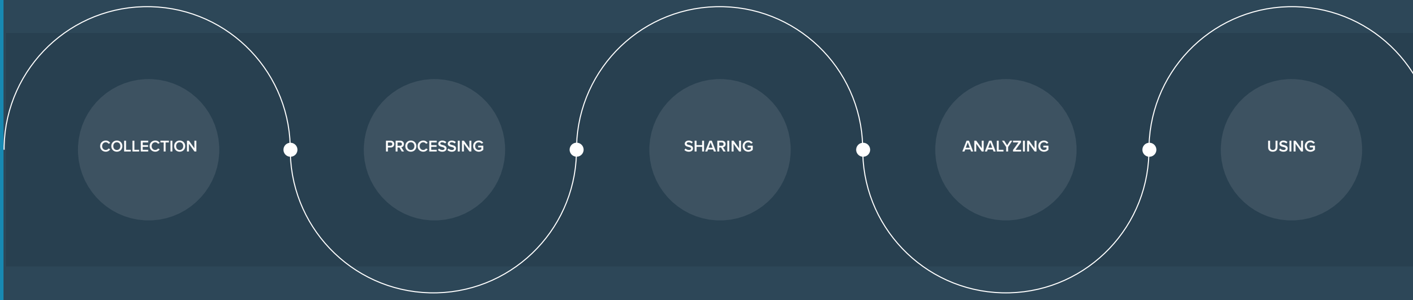


FIGURE 36

Assessing those risks using templates like the 'Data Innovation Risk Assessment Tool'[204] developed by UN Global Pulse is key for determining if and how to share social media data. At the same time, and equally important to assessing the risk of sharing, is the need to assess the public value social media data can generate, and the risk of not sharing. Becoming more systematic about assessing both risk and value is an important task for all data collaborative efforts.

DATA RESPONSIBILITY PRINCIPLES

In addition to a risk/benefit assessment, having a set of principles that can determine if and how social media data can be shared is important for accountability and effectiveness reasons. Obviously, existing principles—such as the Fair Information Practice Principles established by the OECD or the Common Rule principles embedded in various IRBs—remain valuable yet have also proven to be limited when considering new and innovative data practices. To cope with risks and/or opportunities that may not have been on the horizon when previous principles were developed (mainly in the 1970s), data responsibility principles will need to be updated or newly designed. Efforts like the Signal Code[205]—which seeks to apply human rights principles to data during times of crisis—or the Menlo Report[206] that seeks to provide ethical principles guiding Information and Communication Technology research are important steps forward yet often limited to specific contexts and sectors. Similarly the "Handbook on Data Protection in Humanitarian Action,,"[207] recently released by the International Committee of the Red Cross, seeks to provide a set of data processing principles to guide the staff of humanitarian organizations involved in processing personal data as part of humanitarian operations.

DATA RESPONSIBILITY GOVERNANCE PROCESSES AND FUNCTIONS

Principles and assessment frameworks have no value if there is no process in place to apply them or oversee their implementation. Hence the need to include governance processes as a key element to achieve data responsibility. Such processes should ideally be transparent and participatory but also flexible and responsive to different needs and contexts. To accommodate collaborative research using social media data, Facebook, for instance, designed an innovative review process[208] that involved in-house training; different stages of review and the application of evaluation criteria to determine whether to go ahead (or not).

NEW METHODS AND TECHNOLOGIES

Finally, Data Responsibility Frameworks will also need to move beyond written policies and develop tools and technologies that enable flexibility and context specific implementation. For instance, data responsibility decision trees—such as the Center for Democracy & Technology's Digital Decision Tool,[209] which translates principles into a series of questions; or the consideration and experimentation of differential privacy[210]—can be used to limit the privacy risks. Similarly, as concerns about how tools such as public sentiment analysis could be misused by governments may increase, having a transparency report could help allay some concerns: disclosing once or twice a year what data collaboratives have been established.

V. REALIZING THE POTENTIAL OF SOCIAL MEDIA

Intelligence through Data Collaboration

Social Media Data collaboratives clearly hold significant promise as a means of leveraging ***private*** social media data toward ***public*** good. They offer a model for inter-sectoral collaboration across the data value chain, and for applying dispersed expertise and knowledge to untapped stores of social media data silos. Designed and implemented correctly, they offer win-win propositions both to data holders and data recipients. In addition, although certain risks and challenges inevitably exist, these too can be minimized by thoughtful design and implementation, as well as by oversight mechanisms and procedures.

Despite this growing recognition of the potential benefits (and challenges) of social media data collaboratives, these new forms of public-private partnerships remain generally under-utilized as solutions for large, complex public problems. In part, this is no doubt a result of their relative newness: there exists little systematic research or analysis to help us understand the best way to form and utilize data collaboratives. In particular, we have little understanding of how to best "institutionalize" the sharing of data for public good in a responsible manner within corporations as well as their public sector and philanthropic partners.

In this concluding section, we recommend four priorities toward realizing the potential of social media intelligence through data collaborations. Using the preceding case studies and examples, as well as additional research conducted by the GovLab, we have isolated key steps that can help unlock the potential of data collaboratives, and in the process the underlying data itself. We group these steps into four broad categories (recognizing, of course the occasional overlap and interdependence of these categories): stewards; evidence; methods; and movement. Taken together, they can be considered a roadmap for social media companies considering the possibilities of data sharing and the greater use of private data for public good.

STEWARDS

This report makes clear that there is increasing awareness of and experimentation with using social media data toward the public good. However, despite the many case studies discussed here, it is clear that these projects remain more the exception than the rule. Corporations—including but not limited to social media companies—generally fail to recognize realistic opportunities where the data they collect and hold could create public value. In addition, the various concerns discussed above—especially worries over privacy, security, and competitive advantage—can lead to a risk-averse mindset wherein data that could make a difference is not shared with those who could use it.

It is our observation that one of the main reasons for the under-utilization of social media is that social media corporations have rarely institutionalized the function of "data stewardship." Chief data officers and other data-focused roles are understandably focused on how data held by the corporation can benefit businesses' bottom lines. Generally, there is no function with the responsibility for exploring how the company's data could be leveraged for non-corporate purposes, and assessing the balance between risks and rewards of sharing that data. Throughout our research for this report, we identified several data collaborative champions within corporations; yet only a few seemed to have a dedicated role to scale the use of corporate data for public good. Moreover, as discussed above, since data collaboratives often move beyond a simple private to public data-sharing approach—with other actors bringing skills, resources, and data to bear—data stewards can take on a facilitation role to ensure actors successfully work in concert with one another toward common ends.

Given the above, we make the following recommendations:

- Social media corporations should consider themselves, and act as, the standard bearers for a new corporate paradigm of *data stewardship*. This coinage represents a move away from the concept of data as something to be owned, towards stewardship of data as a public good.
- Social media companies should pioneer the role or position of Data Stewards within their organizations. Such individuals (or offices) would be tasked with matching and coordinating the demand and supply of social media data in a trusted way. These new positions could provide models for other companies considering the use of private data for public goals.

- Among their other roles, Data Stewards could help develop new coordinating mechanisms to unlock corporations' supply of social media data sets with potential public interest value. Such mechanisms must include a due process to respond to data requests; a system for filtering or prioritizing certain kinds of information; and a method to ensure that the data being released matches public needs and demands.

EVIDENCE

Many of the above case studies reveal the obstacles posed by a lack of clear metrics. Such metrics are essential to understanding the outcome of specific data collaborative initiatives, and to either improving the design of existing metrics or designing future instances that can build on existing lessons and best practices. Metrics are also critical to helping companies understand the value of their own data, and in particular how sharing can not only benefit the wider public but also increase their own competitiveness (thus addressing some of the concerns discussed in Section IV).

In order to develop a more robust evidence base for data collaborative, we suggest the following steps:

- A more detailed repository of case studies should be established to document impact and practice. Such a repository, which could build on the foundation offered by the case studies presented here, would highlight what works in forming these new partnerships — their value propositions, technical arrangements, legal frameworks, etc. — and strategies for measuring impact.
- It may not be enough simply to strengthen existing metrics; we may also need new metrics. The difficulties in assessing impact of the case studies included here did not just stem from a lack of evidence on impact or progress. Often, there was simply no clear understanding or agreement on what would constitute progress. To improve the effectiveness of social media data collaboratives, and to help move the space toward a more integrated process rather than one-off initiatives, we need clearer indications of value and impact.
- We also need to ascertain a more granular understanding of what data matches what information need by developing data audit methodologies and tools. Different social media platforms are better positioned for different types of data collaboratives. As evidenced by cases like the public sentiment analysis in El

Salvador and flu tracking, more open and easily aggregated social networks like Twitter are well-suited for larger-scale research and population-wide analytical efforts. Social networks like Facebook that both tend to collect more in-depth personal information and, as a result, see their users opt into more robust privacy controls, are better-suited for data collaboratives that are more tightly controlled and targeted toward more specific interventions—such as the Disaster Maps and Zika messaging initiatives. Developing a data-audit tool that matches types of data with types of questions and purposes in an interactive manner may provide for more targeted design and implementation of social media data collaboratives.

METHODS

Data collaboratives are a nascent entity, and the entire field of practice surrounding them is only just beginning to come into view. What many of the case studies above suggest (e.g., research surrounding the use of social media data to predict flooding or to better understand the perspectives of segments of the population to enable informed communications around immunization) is that existing practices and methods may not be sufficient, or may at the very least need radical updating. In order to scale the use of data collaboratives and achieve the full potential of the underlying data, we need a far better understanding, systematization and perhaps institutionalization of the methods surrounding their deployment.

To that end, we suggest the following:

- Mine existing data collaborative experiments (for example, the ones contained in this report) for valuable lessons about what works and what doesn't. The case studies included in this report include but a sampling of the available evidence; a far more robust evidence base can be conducted by researching and storing (e.g., in an online database) the many more examples of data collaboratives from around the world. A detailed analysis of existing practices contained within this evidence base is the best way to build a ground-up and more granular understanding of how best to design a social media data collaborative.
- Lessons and observations taken from this "mining" exercise can then be translated and shared as a toolkit or roadmap for corporations considering sharing data. They would form the basis of a published (and ideally, freely available) "Best Practices" kit that would constantly be updated and improved, based on the latest evidence.

- These types of methods can help shape our understanding of the optimal structure and design of data collaboratives. In addition, we need a better understanding of emerging techniques practiced within data collaboratives themselves, and whose efficacy greatly contributes to the success (or failure) of such projects. Emergent techniques include natural language processing, neural networks, computational social science, network science, sentiment analysis, data-mining and machine-learning. As these disciplines continue to evolve, researchers and practitioners—both inside and outside participating organizations—should seek to determine the optimal frameworks for leveraging these new approaches in the most responsible and effective ways.
- The complexity of many of these techniques means that many organizations (perhaps especially in the humanitarian and civil society fields) may lack the necessary expertise to achieve the potential of data collaboratives. Corporations' sharing data toward the public good need also to consider sharing their expertise in handling data—for example, through training initiatives, educational programs, or informal mentoring setups that can spread the knowledge and skills they have acquired over the course of their data gathering exercises.

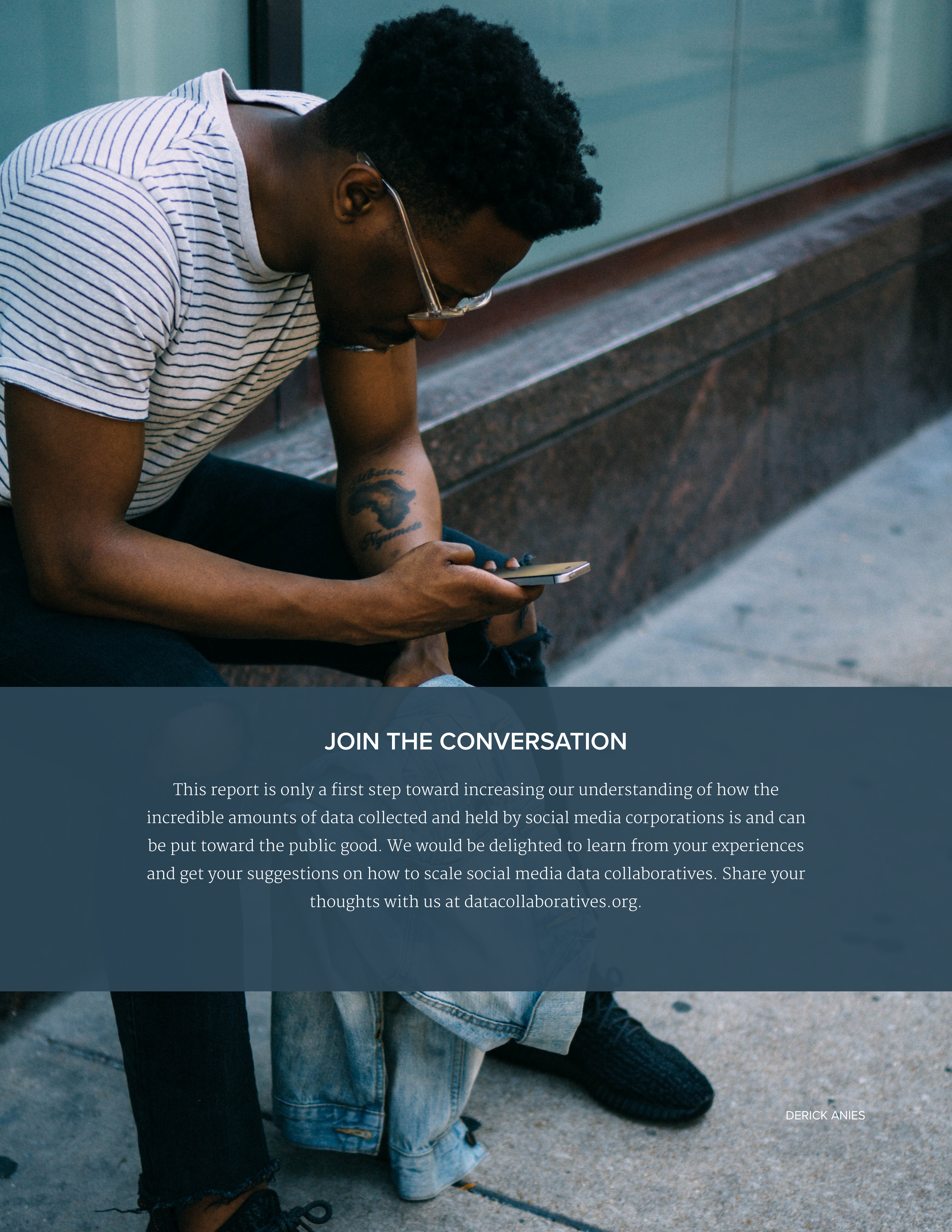
MOVEMENT

By definition, data collaboratives rely on ***collaboration***: the most successful initiatives we studied emerged not in isolation but as part of a wider ecology that encouraged sharing, transparency and accountability. This was the case, for instance, with the Waze Connected Citizens Program, which saw its positive impacts on urban planning directly result from close collaboration between technologists, data scientists and city government officials. Efforts to predict adverse drug reactions using social media data, on the other hand, have been hampered by the lack of a sustaining environment, and this without doubt was one of the reasons for the limited impact in practice, despite the clear viability of the methodological approach.

Our broad recommendation here is that data sharing cannot be thought of in isolation from other activities or environments. Much more thought—and energy—needs to be directed toward developing a network for collaboration, and a platform for sharing. We have several suggestions about what this could mean in practice:

- Bring together various actors from the social media data community—including those that are currently acting as data stewards—at dedicated convenings to share lessons learned; identify pain points and develop common solutions, procedures and practices.

- Corporations could help facilitating such convenings by providing a venue (virtual or physical) where data providers and users can build knowledge in an emerging field and co-create ideas and insights. Such convenings can play a key role in developing a movement around social media data sharing.
- Creating a supporting ecology means interacting with—and nurturing—all actors in that ecology. The practice of data sharing needs to move beyond just data, to include a wider set of skills, expertise and knowledge areas. For example, while social media corporations may possess the most relevant data handling skills, they may not have access to the best minds when it comes to understanding social problems or needs. A true data sharing movement would bring together these various skill sets and apply their combined expertise to the vast stores of available data.
- Data subjects themselves represent another important part of the data ecology. As the data collaboratives and social media intelligence movement continues to take shape, engagement with the populations whose attributes and behaviors often make up the data held by social media companies will be key. Many data subjects are likely to support the use of social media data to solve important public problems, but a more direct conversation between data subjects, data collectors, and data users will better position the field for success.
- In addition to engaging with other skills, there also needs to be broader engagement with other actors whose decisions can play an important—perhaps even determining—role in the success of data collaboratives. Policymakers are perhaps the most important set of actors here, and data collaboratives will rely significantly on the orientations and decisions taken by regulators and politicians. In addition, as seen in many of the above case studies, data collaboratives are often starved for funds, and a wider engagement with the funding community—venture capitalists, private equity players, bankers as well as grant-making organizations—can help formulate justifications and avenues for increased funding.
- Finally, as discussed at length in Section IV (and so we only briefly revisit it here), the success of data collaboratives requires a *cultural reorientation* in the ways companies think about data, ownership, private property and sharing. The foundations of a true data collaborative movement rests on a culture of sharing; for many companies, this will mean a fundamental rethink in the way they operate, interact with their customers, and think of the bottom-line.



DERICK ANIES

JOIN THE CONVERSATION

This report is only a first step toward increasing our understanding of how the incredible amounts of data collected and held by social media corporations is and can be put toward the public good. We would be delighted to learn from your experiences and get your suggestions on how to scale social media data collaboratives. Share your thoughts with us at datacollaboratives.org.

References

1 See datacollaboratives.org

2 Tim Althoff, et al, "Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior," Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017, http://dl.acm.org/citation.cfm?doid=3018661.3018672.

3 Simon Kemp, "Number of Social Media Users Passes 3 Billion with No Signs of Slowing," Contributors, July 8, 2017, https://thenextweb.com/contributors/2017/08/07/number-social-media-users-passes-3-billion-no-signs-slowing/#.tnw_zEiGCS5w.

4 Kaitlin Dewey, "98 Personal Data Points that Facebook Uses to Target Ads to You," Washington Post, August 19, 2016, https://newsroom.fb.com/news/2017/06/two-billion-people-coming-together-on-facebook/.

5 Khalid Saleh, "Social Media Engagement—Statistics and Trends," Invesp blog, https://www.invespcro.com/blog/social-media-engagement/.

6 Cooper Smith, "Social Big Data: The user data collected by each of the World's largest social networks—and what it means," Business Insider, February 2, 2014, http://www.businessinsider.com/social-big-data-the-type-of-data-collected-by-social-networks-2-2014-1.

7 "Twitter Usage Statistics," Internet Live Stats, http://www.internetlivestats.com/twitter-statistics/

8 Ibid.

9 Simon Kemp, "Number of Social Media Users Passes 3 Billion with No Signs of Slowing," Contributors, July 8, 2017, https://thenextweb.com/contributors/2017/08/07/number-social-media-users-passes-3-billion-no-signs-slowing/#.tnw_zEiGCS5w.

10 Bogdan Batrinca and Philip C. Treleaven, "Social Media Analytics: A survey of techniques, tools and Platforms," AI and Society, vol. 30, 1, February 2015, https://link.springer.com/article/10.1007/s00146-014-0549-4.

11 GovLab, "Data Collaboratives: Creating public value by exchanging data," http://datacollaboratives.org/.

12 Iryna Susha, Marijn Janssen and Stefaan Verhulst, "Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy development," Proceedings of the 50th Hawaii International Conference on System Sciences, 2017, http://scholarspace.manoa.hawaii.edu/bitstream/10125/41481/1/paper0332.pdf.

13 U.N. Global Pulse, GSMA, "State of Mobile Data for Social Good," Report Preview, February 2017, https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/02/State-of-Mobile-Data-for-Social-Good-Preview-Feb-2017.pdf.

14 Silja M. Eckartz, Wout J. Hofman, Anne Fleur Van Veenstra, "A Decision Model for Data Sharing," in Electronic Government, https://link.springer.com/chapter/10.1007%2F978-3-662-44426-9_21#enumeration.

15 Stefaan Verhulst, "Corporate Social Responsibility for a Data Age," Stanford Social Innovation Review, February 15, 2017, https://ssir.org/articles/entry/corporate_social_responsibility_for_a_data_age.

16 UNICEF, "Malawi Humanitarian Situation Report," March 11, 2015, https://www.unicef.org/malawi/MLW_floodresponse_humanitariansitrep13.pdf.

17 Sarah Rawson and David Orr, "Japan Helps Feed Flood-Affected Families and School Children in Malawi," World Food Programme, March 6, 2015, https://www.wfp.org/news/news-release/japan-helps-feed-flood-affected-families-and-school-children-malawi.

18 "Half of Malawi Declared Disaster Zone after Flooding," The Guardian, January 14, 2015, https://www.theguardian.com/global-development/2015/jan/14/malawi-flooding-torrential-rain-mozambique.

19 Mark South, Malawi Red Cross Appeals for Continued Flood Operation Support," International Federation of Red Cross and Red Crescent Societies, January 29, 2008, http://www.ifrc.org/ar/news-and-media/news-stories/africa/malawi/malawi-red-cross-appeals-for-continued-flood-operation-support/.

[20] Richard Nield, "Devastation and Disease after Deadly Malawi Floods," Aljazeera, February 25, 2015, http://www.aljazeera.com/indepth/features/2015/02/devastation-disease-deadly-malawi-floods-150225070745817.html.

[21] ITC, "Mapping Party Putting Malawi on the Map," University of Twente, November 2016, https://www.itc.nl/Pub/Events-Conferences/2016/november-2016/Mapping-Party-putting-Malawi-on-the-map.html.

[22] American Red Cross, "Data from Facebook Helps Red Cross Make Better Maps," ARC, November 15, 2016, http://www.redcross.org/news/article/Data-from-Facebook-helps-Red-Cross-make-better-maps&sa=D&ust=1490725501952000&usg=AFQjCNGYtz-T7mrgkial-lSAQ7yCwyAXIA.

[23] Missing Maps website, http://www.missingmaps.org/.

[24] GovLab, "Data Collaboratives: Creating public value by exchanging data," http://datacollaboratives.org/.

[25] Mark Zuckerberg, Facebook post, April 30, 2015, https://www.facebook.com/photo.php?fbid=10102060884238261&set=a.612287952871.2204760.4&type=3&theater.

[26] Mark Zuckerberg, "Buildiing Global Community," Facebook, February 16, 2017, https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634?qid=339663931&mf_story_key=10103508221158471.

[27] Catherine Cheney, "Facebook Introduces Disaster Maps Announces Early Partners," Devex, June 7, 2017, https://www.devex.com/news/facebook-introduces-disaster-maps-announces-early-partners-90427.

[28] Ibid.

[29] Molly Jackman, "Using Data to Help Communities Recover and Rebuild," Facebook Newsroom, June 7, 2017, https://newsroom.fb.com/news/2017/06/using-data-to-help-communities-recover-and-rebuild/.

[30] Matt Petronzio, "Facebook's New 'Disaster Maps' Could Revolutionize Natural Disaster Rescue Efforts," Mashable, June 7, 2017, http://mashable.com/2017/06/07/facebook-disaster-maps-humanitarian-aid/#kCnOfFxDPOq9.

[31] Molly Jackman, "Using Data to Help Communities Recover and Rebuild," Facebook Newsroom, June 7, 2017, https://newsroom.fb.com/news/2017/06/using-data-to-help-communities-recover-and-rebuild/.

[32] https://www.google.org/crisismap/weather_and_events.

[33] GovLab interview with Paige Maas, Facebook, August 29, 2017.

[34] Michael Sheetz, "Facebook Says Its New Disaster Relief Tool Tool Will Keep Users' Data Private," CNBC, June 8, 2017, https://www.cnbc.com/2017/06/08/facebook-disaster-relief-tool-company-promises-privacy-and-safety.html.

[35] Catherine Cheney, "Facebook Introduces Disaster Maps Announces Early Partners," Devex, June 7, 2017, https://www.devex.com/news/facebook-introduces-disaster-maps-announces-early-partners-90427.

[36] Molly Jackman, "Using Data to Help Communities Recover and Rebuild," Facebook Newsroom, June 7, 2017, https://newsroom.fb.com/news/2017/06/using-data-to-help-communities-recover-and-rebuild/.

[37] GovLab interview with Paige Maas, Facebook, August 29, 2017.

[38] Matt Petronzio, "Facebook's New 'Disaster Maps' Could Revolutionize Natural Disaster Rescue Efforts," Mashable, June 7, 2017, http://mashable.com/2017/06/07/facebook-disaster-maps-humanitarian-aid/#S1m7ED6BiOqi.

[39] Paige Maas, et al., "Facebook Disaster Maps: Methodology," Facebook, June 7, 2017, https://research.fb.com/facebook-disaster-maps-methodology/,

[40] GovLab interview with Paige Maas, Facebook, August 29, 2017.

[41] https://www.devex.com/news/facebook-introduces-disaster-maps-announces-early-partners-90427

[42] Ibid.

[43] Lely Djuhari, "Tracking Anti-vaccination Sentiment in Eastern European Social Media Networks," UNICEF Media Center, April 22, 2013, https://www.unicef.org/eca/media_24017.html.

44 UN Global Pulse, "Understanding Immunisation Awareness and Sentiment through Social and Mainstream Media," Global Pulse Project Series no. 19, 2015, http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Perception_Immunisation_2014_0.pdf.

45 https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/

46 Ibid.

47 Tobias Tiecke, "Open Population Datasets and Open Challenges," Facebook posts, November 15, 2016, https://code.facebook.com/posts/596471193873876/open-population-datasets-and-open-challenges/.

48 Andreas Gros and Tobias Tiecke, "Connecting the World with Better Maps," Facebook posts, February 21, 2016, https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/.

49 Facebook Connectivity Lab and Center for International Earth Science Information Network—CIESIN—Columbia University, "High Resolution Settlement Layer (HRSL), 2016, https://ciesin.columbia.edu/data/hrsl/.

50 Facebook, "Connecting the World with Better Maps: Data-assisted Population Distribution Mapping," https://fbnewsroomus.files.wordpress.com/2016/02/population_density_final_mj2_ym_tt2113.pdf.

51 Ibid.

52 Ibid.

53 Robert Chen, "Working with Facebook to Create Better Population Maps," State of the Planet Earth Institute, Columbia University Blog, February 22, 2016, http://blogs.ei.columbia.edu/2016/02/22/working-with-facebook-to-create-better-population-maps/.

54 http://www.redcross.org/news/article/Data-from-Facebook-helps-Red-Cross-make-better-maps

55 Talip Kilic, et al., "A First Look at Facebook's High Resolution Facebook Maps" World Bank, The Data Blog, November 18, 2016, https://blogs.worldbank.org/opendata/first-look-facebook-s-high-resolution-population-maps

56 Yelp Dataset Challenge, https://www.yelp.com/dataset_challenge.

57 Ibid.

58 Scott Clark, "Yelp Dataset Challenge Winners and Round Two Now Live," Yelp Engineering Blog, October 2, 2013, https://engineeringblog.yelp.com/2013/10/yelp-dataset-challenge-winners-round-two-now-live.html.

59 Yelp Dataset Challenge, https://www.yelp.com/dataset_challenge.

60 Ibid.

61 Felix W., "On the Efficiency of Social Recommender Networks," Yelp, https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_NetworkEfficiency.pdf.

62 Jack Linshi, "Personalizing Yelp Star Ratings: A semantic topic modelling approach," Yelp, https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_PersonalizingRatings.pdf.

63 Demitrios Kotzias, et al., "From Group to Individual Labels Using Deep Features," Knowledge Discovery and Data Mining, 2015, http://mdenil.com/media/papers/2015-deep-multi-instance-learning.pdf.

64 Abhinav Maurya, et al., "Semantic Scan: Detecting subtle, spatially localized events in text streams," February 13, 2016, https://arxiv.org/abs/1602.04393.

65 Chen Li and Jin Zhang, "Prediction of Yelp Star Rating Using Sentiment Analysis," CS 229 Project Final Report, http://cs229.stanford.edu/proj2014/Chen%20Li,%20Jin%20Zhang,%20Prediction%20of%20Yelp%20Review%20Star%20Rating%20using%20Sentiment%20Analysis.pdf.

66 Patrick Bechon, Leo Grimaldi and Yacine Merouchi, "CS 229 Project: Improving on Yelp reviews using NLP and Bayesian Scoring," CS 229 Project, http://cs229.stanford.edu/proj2011/BechonGrimaldiMerouchi-ImprovingYelpReviews.pdf.

67 See e.g., Google Scholar search for "Yelp Dataset," https://scholar.google.com/scholar?q=citation%3A+Yelp+Dataset&btnG=&hl=en&as_sdt=0%2C5.

68 Julian McAuley and Jure Leskovec, "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text," http://i.stanford.edu/~julian/pdfs/recsys_extended.pdf.

69 Berkley School of Information, “Students Data Analysis Uncovers Hidden Trends in Yelp Reviews,” October 4, 2013, https://www.ischool.berkeley.edu/news/2013/students-data-analysis-uncovers-hidden-trends-yelp-reviews.

70 William Powers, “Fueling the Horse Race of Ideas in the 2016 Election,” Laboratory for Social Machines, October 29, 2015, http://socialmachines.media.mit.edu/2015/10/29/fueling-the-horse-race-of-ideas-3/.

71 Ibid.

72 Peter Hamby, “Did Twitter Kill the Boys on the Bus? Searching for a better way to cover a campaign,” Joan Shorenstein Center on the Press, Politics and Public Policy, Discussion Paper Series, September 2013, https://shorensteincenter.org/wp-content/uploads/2013/08/d80_hamby.pdf.

73 Chris O’Shea, “Study: More journalists use Twitter,” Adweek, June 12, 2013, http://www.adweek.com/digital/study-more-journalists-use-twitter/.

74 Jim Edwards, “Leaked Twitter API Data Shows the Number of Tweets Is in Serious Decline,” Business Insider, February 2, 2016, http://www.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2.

75 MIT Media Lab, Laboratory for Social Machines, https://socialmachines.media.mit.edu/.

76 MIT Media Lab, Laboratory for Social Machines, “About,” https://socialmachines.media.mit.edu/about/.

77 MIT Media Lab, “The Electome: Measuring responsiveness in the 2016 election,” www.media.mit.edu/projects/the-electome-measuring-responsiveness-in-the-2016-election/overview/.

78 The Electome, “Enter the Electome: Fueling the horse race of ideas in the 2016 election,” http://www.electome.org/.

79 The Electome, “The Electome Social Pulse: How did twitter watch the debates?” http://www.electome.org/#debates.

80 The Electome, “Dashboard,” http://dashboard.electome.org/.

81 Peter Dizikes, “How Twitter Explains the 2016 Election,” MIT News, September 26, 2016, http://news.mit.edu/2016/how-twitter-explains-the-2016-election-0926.

82 John West, “Immigration Is Dominating the Election Conversation on Twitter,” Washington Post, September 5, 2016, https://www.washingtonpost.com/news/post-politics/wp/2016/09/05/immigration-is-dominating-the-election-conversation-on-twitter/.

83 Bloomberg, “Missing from 2016 Race: Sense of urgency over U.S. budget gaps,” May 2, 2016, Bipartisan Policy Center, https://bipartisanpolicy.org/article/missing-from-2016-race-sense-of-urgency-over-u-s-budget-gaps/.

84 Natalie Andrews, “2016 Front-runners Top List of Election Influencers on Social Media,” Wall Street Journal, March 1, 2016, https://blogs.wsj.com/washwire/2016/03/01/2016-front-runners-top-list-of-election-influencers-on-social-media/.

85 Kate Stohr, “Warning: This election contains language some people may find offensive,” Splinter News, May 25, 2016, http://splinternews.com/warning-this-election-contains-language-some-people-ma-1793857034

86 Alex Thompson, “Parallel Narratives: Clinton and Trump Supporters really don’t listen to each other on Twitter,” Vice News, December 8, 2016, https://news.vice.com/story/journalists-and-trump-voters-live-in-separate-online-bubbles-mit-analysis-shows?utm_source=vicenewsfb.

87 MIT Media Lab, “CNN Politics Data,” http://socialmachines.media.mit.edu/wp-content/uploads/sites/27/2016/07/cnn_politics_1.pdf.

88 LSM Team, MIT Media Lab, “Electome Articles,” February 20, 2017, https://socialmachines.media.mit.edu/2017/02/20/electome-articles/.

89 MIT Media Lab, “The Electome: Measuring responsiveness in the 2016 election,” www.media.mit.edu/projects/the-electome-measuring-responsiveness-in-the-2016-election/overview/

90 Alex Thompson, “Parallel Narratives: Clinton and Trump Supporters really don’t listen to each other on Twitter,” Vice News, December 8, 2016, https://news.vice.com/story/journalists-and-trump-voters-live-in-separate-online-bubbles-mit-analysis-shows?utm_source=vicenewsfb.

91 Shannon Greenwood, et al., “Social Media Update 2016: Facebook usage and engagement is on the rise while adoption of other platforms holds steady,” Pew Research Center Internet and Technology Report, November 11, 2016, http://www.pewinternet.org/2016/11/11/social-media-update-2016/#fn-17239-1.

92 Scott Clement, “The Political Debate on Twitter Focuses on Much Different Issues than Americans at Large,” Washington Post, March 13, 2016, https://www.washingtonpost.com/news/the-fix/wp/2016/03/13/the-political-debate-on-twitter-focuses-on-much-different-issues-than-americans-at-large/.

93 Deepak Agarwal, “Big Thinkers, Big Data, Big Opportunity: Announcing the LinkedIn economic graph challenge,” LinkedIn Blog, October 14, 2014, https://blog.linkedin.com/2014/10/14/big-thinkers-big-data-big-opportunity-announcing-the-linkedin-economic-graph-challenge

94 Igor Perisic, “Announcing Winners of the LinkedIn Economic Graph Challenge,” LinkedIn Blog, May 11, 2015, https://blog.linkedin.com/2015/05/11/announcing-winners-of-the-linkedin-economic-graph-challenge.

95 Ibid.

96 Deepak Agarwal, “Announcing the Economic Graph Research Program,” LinkedIn Engineering Blog, March 15, 2017, https://engineering.linkedin.com/blog/2017/03/announcing-the-economic-graph-research-program.

97 Pablo Chavez, “How LinkedIn’s Economic Graph Is Helping to Close the Skills Gap,” LinkedIn Blog, September 9, 2016, https://blog.linkedin.com/2016/09/09/how-linkedin_s-economic-graph-is-helping-to-close-the-skills-gap.

98 Deepak Agarwal, “Announcing the Economic Graph Research Program,” LinkedIn Engineering Blog, March 15, 2017, https://engineering.linkedin.com/blog/2017/03/announcing-the-economic-graph-research-program.

99 Pablo Chavez, “How LinkedIn’s Economic Graph Is Helping to Close the Skills Gap,” LinkedIn Blog, September 9, 2016, https://blog.linkedin.com/2016/09/09/how-linkedin_s-economic-graph-is-helping-to-close-the-skills-gap

100 Neil Irwin, “How to Become a CEO? The quickest path is a winding one,” New York Times, September 9, 2016, www.nytimes.com/2016/09/11/upshot/how-to-become-a-ceo-the-quickest-path-is-a-winding-one.html

101 Kate Davidson, “The ‘Soft Skills’ Employers Are Looking For,” Wall Street Journal, August 30, 2016, https://blogs.wsj.com/economics/2016/08/30/the-soft-skills-employers-are-looking-for/.

102 Hsuanwei Michelle Chen, et al, “Exploring Government Uses of Social Media through Twitter Sentiment Analysis,” Journal of Digital Information Managemernt, June 10, 2016, http://dline.info/fpaper/jdim/v14i5/jdimv14i5_2.pdf.

103 Facebook, OECD, and the World Bank, “Future of Business Survey,” https://eu.futureofbusinesssurvey.org/manager/Storyboard/RHViewStoryBoard.aspx

104 Tim Herzog and Joshua Wimpey, “Facebook, the OECD and the World Bank Have a New Way to Survey Businesses,” The World Bank Blog, September 29, 2016, https://blogs.worldbank.org/opendata/facebook-oecd-world-bank-have-new-way-survey-businesses.

105 Facebook, OECD, and the World Bank, “Future of Business Survey,” https://eu.futureofbusinesssurvey.org/manager/Storyboard/RHViewStoryBoard.aspx

106 Available for download at: https://eu.futureofbusinesssurvey.org/manager/Storyboard/RHViewStoryBoard.aspx

107 Ibid.

108 Facebook, OECD, and the World Bank, “Future of Business Survey,” http://bit.ly/2gR7fwp.

109 Google, “About the Connected Citizens Program,” https://support.google.com/waze/partners/answer/6372611?hl=en.

110 Google, “The Waze Partners Portal and Forum,” https://support.google.com/waze/partners/answer/7239092?hl=en&ref_topic=7237973.

111 Google, “Waze Connected Citizens Program Case Studies,” https://docs.google.com/gview?url=https%3A%2F%2Fs3-eu-west-1.amazonaws.com%2Fwaze-partner-assets%2FCCPCaseStudies.pdf.

112 Google, "About the Connected Citizens Program," https://support.google.com/waze/partners/answer/6372611?hl=en.

113 Google, "Waze Connected Citizens Program Case Studies," https://docs.google.com/gview?url=https%3A%2F%2Fs3-eu-west-1.amazonaws.com%2Fwaze-partner-assets%2FCCPCaseStudies.pdf.

114 Interview with Jascha Franklin-Hodge, CIO, City of Boston, August 24, 2017.

115 Google, "Waze Connected Citizens Program Case Studies," https://docs.google.com/gview?url=https%3A%2F%2Fs3-eu-west-1.amazonaws.com%2Fwaze-partner-assets%2FCCPCaseStudies.pdf.

116. Interview with Jascha Franklin-Hodge, CIO, City of Boston, August 24, 2017.

117. Google, "About the Connected Citizens Program," https://support.google.com/waze/partners/answer/6372611?hl=en.

118 Interview with Jascha Franklin-Hodge, CIO, City of Boston, August 24, 2017.

119 Google, "Waze Connected Citizens Program Case Studies," https://docs.google.com/gview?url=https%3A%2F%2Fs3-eu-west-1.amazonaws.com%2Fwaze-partner-assets%2FCCPCaseStudies.pdf.

120 Elizabeth Weise, "Waze and Other Traffic Dodging Apps Prompt Cities to Game the Algorithms," USA Today, March 6, 2017, https://www.usatoday.com/story/tech/news/2017/03/06/mapping-software-routing-waze-google-traffic-calming-algorithmsi/98588980/.

121 Interview with Jascha Franklin-Hodge, CIO, City of Boston, August 24, 2017.

122 https://www.wired.com/2017/02/ubers-coughing-data-nyc-fix-commute/

123 Interview with Jascha Franklin-Hodge, CIO, City of Boston, August 24, 2017.

124 UNICEF Innovation, "How Facebook Statuses Informed the Zika Response in Brazil," UNICEF Blog, February 10, 2017, https://blogs.unicef.org/innovation/facebook-statuses-informed-zika-response-brazil/.

125 Ibid.

126 Insights for Impact, "A Case Study: Data and Social Media Can Lead to Healthier Lives," UNICEF and Facebook, http://neo-assets.s3.amazonaws.com/news/FB-UNICEF-Big.png.

127 Ibid.

128 Ibid.

129 Interview with Drew Bernard, ActionSprout, August 28, 2017.

130 Insights for Impact, "A Case Study: Data and Social Media Can Lead to Healthier Lives," UNICEF and Facebook, http://neo-assets.s3.amazonaws.com/news/FB-UNICEF-Big.png.

131 Interview with Drew Bernard, ActionSprout, August 28, 2017.

132 UNICEF Innovation, "How Facebook Statuses Informed the Zika Response in Brazil," UNICEF Blog, February 10, 2017, https://blogs.unicef.org/innovation/facebook-statuses-informed-zika-response-brazil/.

133 National Center for Health Statistics, "Influenza," https://www.cdc.gov/nchs/fastats/flu.htm.

134 Centers for Disease Control and Prevention, "Overview of Influenza Surveillance in the United States," https://www.cdc.gov/flu/weekly/overview.htm.

135 Matthew Biggerstaff, et al., Results from the Centers for Disease Control and Prevention's Predict the 2013-14 Influenza Season Challenge," BMC Infectious Diseases, vol. 16 (2016), https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-016-1669-x.

136 Michael J. Paul, et al., "Twitter Improves Influenza Forecasting," Plos, October 28, 2014, http://currents.plos.org/outbreaks/article/twitter-improves-influenza-forecasting/.

137 Michael J. Paul and Mark Dredze, "You Are What You Tweet: Analyzing Twitter for public health," Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2880/3264.

138 Ibid.

139 Michael J. Paul, et al., "Twitter Improves Influenza Forecasting," Plos, October 28, 2014, http://currents.plos.org/outbreaks/article/twitter-improves-influenza-forecasting/.

140 Ibid.

141 Centers for Disease Control and Prevention, "CDC Competition Encourages Use of Social Media to Predict Flu," November 25, 2013, https://www.cdc.gov/flu/news/predict-flu-challenge.htm.

142 Centers for Disease Control and Prevention, "Flu Activity Forecasting Website Launched," January 19, 2016, https://www.cdc.gov/flu/news/flu-forecast-website-launched.htm.

143 Matthew Biggerstaff, et al., Results from the Centers for Disease Control and Prevention's Predict the 2013-14 Influenza Season Challenge," BMC Infectious Diseases, vol. 16 (2016), https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-016-1669-x.

144 https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-016-1669-x

145 David Lazer, et al., "The Parable of Google Flu: Traps in big data analysis," Science, vol. 343, 6176, (March 14, 2014), http://science.sciencemag.org/content/343/6176/1203.full.

146 GovLab interview with Nataliya Tkachenko, August 25, 2017.

147 Nataliya Tkachenko, et al., "Predicting Floods with Flickr Tags," Plos One, February 24, 2017, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0172870.

148 Ibid.

149 https://www.floodtags.com/

150 GovLab interview with Nataliya Tkachenko, August 25, 2017.

151 Tobias Preis, et al., "Quantifying the Digital Traces of Hurricane Sandy on Flickr," Scientific Reports, vol. 3, November 5, 2013, https://www.nature.com/articles/srep03141.

152 Nataliya Tkachenko, et al., "Predicting Floods with Flickr Tags," Plos One, February 24, 2017, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0172870.

153 Ibid.

154 GovLab interview with Nataliya Tkachenko, August 25, 2017.

155 Ibid.

156 Daphne E. Smith Marsh, "Adverse Drug Reactions," Merck Manual: Professional edition, http://www.merckmanuals.com/professional/clinical-pharmacology/adverse-drug-reactions/adverse-drug-reactions.

157 A. Nikfarjam and G.H. Gonzales, "Pattern Mining for Extraction of Mentions of Adverse Drug Reactions from User Comments," AMIA Annual Symposium Proceedings, October 22, 2011, https://www.ncbi.nlm.nih.gov/pubmed/22195162.

158 Ronen Feldman, et al., "Utilizing Text Mining on Online Medical Forums to Predict Label Change Due to Adverse Drug Reactions," Youtube Video, October 9, 2015, https://www.youtube.com/watch?v=0Jnh3q6Er34.

159 Isabel Segura-Bedmar, et al., "Detecting Drugs and Adverse Events from Spanish Health Social Media Streams," http://www.aclweb.org/anthology/W14-1117.

160 Ronen Feldman, et al., "Utilizing Text Mining on Online Medical Forums to Predict Label change Due to Adverse Drug Reactions," Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, http://dl.acm.org/citation.cfm?id=2788608.

161 A. Nikfarjam and G.H. Gonzales, "Pattern Mining for Extraction of Mentions of Adverse Drug Reactions from User Comments," AMIA Annual Symposium Proceedings, October 22, 2011, https://www.ncbi.nlm.nih.gov/pubmed/22195162.

162 Isabel Segura-Bedmar, et al., "Detecting Drugs and Adverse Events from Spanish Health Social Media Streams," http://www.aclweb.org/anthology/W14-1117.

163 Ibid.

164 Ibid.

165 Ronen Feldman, et al., "Utilizing Text Mining on Online Medical Forums to Predict Label change Due to Adverse Drug Reactions," Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, http://dl.acm.org/citation.cfm?id=2788608.

166 A. Nikfarjam and G.H. Gonzales, "Pattern Mining for Extraction of Mentions of Adverse Drug Reactions from User Comments," AMIA Annual Symposium Proceedings, October 22, 2011, https://www.ncbi.nlm.nih.gov/pubmed/22195162.

167 Interview with Kate Dale, Head of Brand and Digital Strategy, August 23, 2017.

168 "Case Study: How 'This Girl Can' got 1.6 million women exercising," Campaign, http://www.campaignlive.co.uk/article/case-study-this-girl-can-16-million-women-exercising/1394836

169 Crimson Hexagon and Sport England, "#ThisGirlCan: Social Insights Challenge Women to Rethink Traditional Attitudes to Sport," 2015, http://pages.crimsonhexagon.com/WC2016-01-14-CS-ThisGirlCan_Registration.html.

170 Ibid.

171 Crimson Hexagon and Sport England, "#ThisGirlCan: Social Insights Challenge Women to Rethink Traditional Attitudes to Sport", 2015.

172 "Case Study: How 'This Girl Can' got 1.6 million women exercising," Campaign, http://www.campaignlive.co.uk/article/case-study-this-girl-can-16-million-women-exercising/1394836

173 Interview with Kate Dale, Head of Brand and Digital Strategy, Sport England, August 23, 2017.

174 "Case Study: How 'This Girl Can' got 1.6 million women exercising," Campaign, http://www.campaignlive.co.uk/article/case-study-this-girl-can-16-million-women-exercising/1394836

175 Interview with Kate Dale, Head of Brand and Digital Strategy, Sport England, August 23, 2017.

176 Ibid.

177 Crimson Hexagon and Sport England, "#ThisGirlCan: Social Insights Challenge Women to Rethink Traditional Attitudes to Sport," 2015, http://pages.crimsonhexagon.com/WC2016-01-14-CS-ThisGirlCan_Registration.html.

178 "Case Study: How 'This Girl Can' got 1.6 million women exercising," Campaign, http://www.campaignlive.co.uk/article/case-study-this-girl-can-16-million-women-exercising/1394836.

179 Ibid.

180 Interview with Kate Dale, Head of Brand and Digital Strategy, Sport England, August 23, 2017.

181 Ibid.

182 Ibid.

183 "Case Study: How 'This Girl Can' got 1.6 million women exercising," Campaign, http://www.campaignlive.co.uk/article/case-study-this-girl-can-16-million-women-exercising/1394836.

184 Ibid.

185 Interview with Kate Dale, Head of Brand and Digital Strategy, Sport England, August 23, 2017.

186 Ibid.

187 "Case Study: How 'This Girl Can' got 1.6 million women exercising," Campaign, http://www.campaignlive.co.uk/article/case-study-this-girl-can-16-million-women-exercising/1394836.

188 UN Global Pulse, "Using Twitter Data to Analyse Public Sentiment on Fuel Subsidy Policy Reform in El Salvador," Global Pulse Project Series, no.13, 2015, http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_ElSalvador_Fuel_2015_0.pdf.

189 Ibid.

190 Ibid.

191 Interview with Miguel Luengo-Oroz, August 28, 2017.

192 UN Global Pulse, "Using Twitter Data to Analyse Public Sentiment on Fuel Subsidy Policy Reform in El Salvador," Global Pulse Project Series, no.13, 2015, http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_ElSalvador_Fuel_2015_0.pdf.

193 René Clausen Nielsen, et al., "Social Media Monitoring of Discrimination and HIV Testing in Brazil, 2014-2015," AIDS and Behavior, July 2017. https://link.springer.com/article/10.1007%2Fs10461-017-1753-2.

194 Interview with Miguel Luengo-Oroz, August 28, 2017.

195 UN Global Pulse, "Using Twitter to Measure Global Engagement on Climate Change," Global Pulse Project Series, no.7, 2015, http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Climate_Monitor_2015_0.pdf.

196 UN Global Pulse, "How the World Tweets: Climate change," www.unglobalpulse.net/climate.

197 UN Global Pulse, "Using Twitter to Measure Global Engagement on Climate Change," Global Pulse Project Series, no.7, 2015, http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Climate_Monitor_2015_0.pdf.

198 Ibid.

199 Ibid.

200 Ibid.

201 Interview with Miguel Luengo-Oroz, August 28, 2017.

202 Freddy de Meersman, et al., "Assessing the Quality of Mobile Phone Data as a Source of Statistics," European Conference on Quality in Official Statistics, 2016, https://ec.europa.eu/eurostat/cros/system/files/assessing_the_quality_of_mobile_phone_data_as_a_source_of_statistics_q2016.pdf.

203 Randy Bean, "MasterCard's Big Data for Good Initiative: Data Philanthropy on the Front Lines," Forbes, August 7, 2017, https://www.forbes.com/sites/ciocentral/2017/08/07/mastercards-big-data-for-good-initiative-data-philanthropy-on-the-front-lines/#2911106d20dc.

204 UN Global Pulse, "Data Innovation Risk Assessment Tool, 2016, http://unglobalpulse.org/privacy/tools.

205 Signal Code website, https://signalcode.org/.

206 The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research: https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/

207 International Committee of the Red Cross, "Handbook on Data Protection in Humanitarian Action," August 23, 2017, https://www.icrc.org/en/publication/handbook-data-protection-humanitarian-action.

208 Molly Jackman and Laurie Kanerva, "Evolving the IRB: Buildiing robust review for industry research," Washington and Lee Law Review Online, vol. 72, 3, (2016), http://scholarlycommons.law.wlu.edu/cgi/viewcontent.cgi?article=1042&context=wlulr-online.

209 Natasha Duarte, "Digital Decisions Tool," August 8, 2017, Center for Democracy and Technology Blog, https://cdt.org/blog/digital-decisions-tool/.

210 Matthew Greenin Apple, "What Is Differential Privacy?" https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/.

THE POTENTIAL OF SOCIAL MEDIA INTELLIGENCE
TO IMPROVE PEOPLE'S LIVES

Social Media Data for Good

GOVLAB

facebook