On Building Better Mousetraps and Understanding the Human Condition: Reflections on Big Data in the Social Sciences

By
JIMMY LIN

Over the past few years, we have seen the emergence of "big data": disruptive technologies that have transformed commerce, science, and many aspects of society. Despite the tremendous enthusiasm for big data, there is no shortage of detractors. This article argues that many criticisms stem from a fundamental confusion over goals: whether the desired outcome of big data use is "better science" or "better engineering." Critics point to the rejection of traditional data collection and analysis methods, confusion between correlation and causation, and an indifference to models with explanatory power. From the perspective of advancing social science, these are valid reservations. I contend, however, that if the end goal of big data use is to engineer computational artifacts that are more effective according to well-defined metrics, then whatever improves those metrics should be exploited without prejudice. Sound scientific reasoning, while helpful, is not necessary to improve engineering. Understanding the distinction between science and engineering resolves many of the apparent controversies surrounding big data and helps to clarify the criteria by which contributions should be assessed.

Keywords: big data; computational social science; machine learning; data mining; log analysis

Over the past few years, we have seen the emergence of "big data": disruptive technologies that have transformed commerce, science, and many aspects of society. In the commercial sphere, Google led the way with the development of a large-scale computing infrastructure for storing and analyzing the web and the behavior of hundreds of millions of

Jimmy Lin is an associate professor in the College of Information Studies (The iSchool) and the Institute for Advanced Computer Studies (UMIACS) at the University of Maryland. His research focuses on large-scale distributed algorithms and infrastructure for data analytics. Between 2010 and 2012, he spent an extended sabbatical at Twitter working on services to surface relevant content to users and analytics infrastructure to support data science.

DOI: 10.1177/0002716215569174

users (Dean and Ghemawat 2004). Through investments in open-source software by Yahoo, these innovations rapidly spread to other Internet companies such as Facebook, Twitter, LinkedIn, Amazon, eBay, and countless startups. Big data have attracted the attention of mature technology companies such as IBM, Microsoft, Oracle, and Intel; and beyond the technology sector, organizations ranging from WalMart to J.P. Morgan Chase have all hopped on the big data bandwagon.

In a different realm, researchers in the physical sciences, particularly high-energy physics, have also seen an explosion in the amount of data generated from scientific instruments as well as demand for computing power to analyze those data (Becla and Wang 2005). Today, scientists speak of data-driven research as the "fourth paradigm" (Hey, Tansley, and Tolle 2009), complementing theory, experiments, and simulations. The hunt for the Higgs Boson, for example, was a data-driven endeavor guided by theoretical models.

Big data have similarly been a boon for social scientists, leading to the rise of computational social science (Lazer et al. 2009). Many branches of the social sciences study human individual and group behavior, and thus records of human activity, ranging from social interactions to political preferences, are invaluable for hypothesis generation and empirical validation of constructs and theories. For much of the twentieth century, data collection has been slow and tedious. In the 1970s, Zachary painstakingly documented the network of friendships between thirty-four members of a karate club at a U.S. university, leading to seminal work on information flow and group conflict (Zachary 1977). A couple of years ago, researchers at Facebook and their collaborators published papers (Ugander et al. 2011; Backstrom et al. 2012) describing analyses of the company's massive worldwide social network, totaling 721 million users at the time. Today, it is possible to examine human activities at scales undreamt of a generation ago, and these digital footprints have the potential to help social scientists better understand the complexities of human behavior—for example, how individuals form and maintain social ties and the dynamics of influence and power. In this context, there has been substantial discussion about the merits of massively data-driven approaches in the social sciences. Is more always better? What about data quality and data access? How does big data reshape the nature of knowledge and the activities of science? Some have taken a generally negative view of this development, such as the "provocations" of boyd and Crawford (2011). Lazer et al. (2014) caution against "big data hubris," which is the tendency to assume that big data are a substitute for traditional data collection and analysis, and that we can ignore foundational issues such as construct validity. Other "symptoms" include confusing correlation with causation and an indifference to models with explanatory power.

NOTE: This work was supported in part by the U.S. National Science Foundation under award IIS-1218043. Any opinions, findings, conclusions, or recommendations expressed are those of the author and do not necessarily reflect the views of the sponsor. I'd like to thank Mark Dredze, Louiqa Raschid, and Ben Shneiderman for helpful comments on previous drafts. Finally, I'm grateful to Esther for her loving support and dedicate this work to Joshua and Jacob.

This article comes to the defense of big data and presents personal reflections on its implications. My central thesis is that much of the controversy stems from a fundamental confusion about the purpose of big data. Is the goal "better science"—to reveal insights about the human condition? Or is it about "better engineering"—to build better mousetraps? I contend that if the objective is the latter—to produce computational artifacts that are more effective according to well-defined metrics—then whatever improves those metrics should be exploited without prejudice. Sound scientific reasoning, while helpful, is not necessary to improve engineering. Once we understand the distinction between the fundamentally different goals of what I call "better science" or "better engineering," it becomes clear that many criticisms of big data simply miss the point. This article focuses on problems involving the prediction of human behavior, and in this context, I claim that understanding is not necessary for prediction.

As a simple analogy, consider online matchmaking services—their data have the potential to help sociologists understand human attraction on a large scale. However, those services exist to make a profit, and so they are primarily concerned with issues such as revenue and subscriber growth. Whether any of their matchmaking algorithms are based on our current (academic) understanding of attraction is largely irrelevant (although we hope that the literature provides helpful insights). Similarly, it seems unreasonable to criticize an online matchmaking service for its unwillingness to share member profile data to help social scientists advance their research agenda.

Before proceeding further, it is important to carefully circumscribe the scope of my claims. In this article, I limit my commentary primarily to social media and the activities of online users. For the most part, these involve "low-stakes" interactions with negligible health risks and financial impact. There are, however, many more applications of big data, for example, in electronic medical records, networks of sensors (e.g., "smart cities"), online education, and so on. Although some of the discussions here are applicable to scenarios such as public health interventions or learning analytics, there are domain-specific subtleties that defy broad generalizations. Furthermore, some of these applications involve higher-stakes interactions that require much more care before interventions are deployed. In-depth discussion of these application areas is beyond the scope of this article, however.

Background and Context

What is different about big data?

In the business context, big data are the (somewhat obvious) idea that an organization should retain data that result from carrying out its mission and exploit those data to generate insights for better decision-making. Also known as business intelligence, among other monikers, its origins date back several decades. In this sense, the big data hype is simply a rebranding of what many organizations have been doing for a long time. Today, these activities are known as *data science* and those who practice it are known as *data scientists*.

Examined more closely, however, there are three major trends that distinguish insight-generation activities today from, say, the 1990s. First, we have seen a tremendous explosion in the sheer amount of data—orders of magnitude increase. In the past, enterprises have typically focused on gathering data that are obviously valuable, such as business objects representing customers, items in catalogs, purchases, contracts, and so on. Today, in addition to such data, organizations also gather behavioral data from users. In the online setting, these include web pages that users visit and links that they click on, among others. The advent of social media and user-generated content, and the resulting interest in encouraging such interactions, further adds to the amount of data that is generated and collected. These are precisely the types of data that are valuable for computational social science.

Second, we see increasing sophistication in the types of analyses that organizations perform on their vast data stores. Traditionally, most information needs fell under what is known as online analytical processing (OLAP). Common tasks include creating joined views, followed by filtering, aggregation, or cube materialization—an example might be "show me the number of widgets sold in the northeast over the past six months to female customers." We can characterize these activities as descriptive analytics, generating reports that an executive might consume. Today, data scientists are also interested in predictive analytics, which often involves building machine-learned models that can predict user behavior; for example, "what types of targeted ads can attract female customers to purchase this widget?" These models are then operationalized into *data products* that apply some sort of intervention (e.g., a recommender system) to hopefully affect user behavior.

Finally, open-source software is playing an increasingly important role in today's ecosystem. A decade ago, there was no credible open-source, distributed data analytics platform capable of handling large data volumes. Today, the open-source Hadoop platform, which began as an implementation of MapReduce (Dean and Ghemawat 2004), lies at the center of an ecosystem for large-scale data analytics, and is surrounded by complementary systems such as HBase, Pig, Hive, Spark, Giraph, and many others. Hadoop's importance has been validated by its adoption in countless startups and mature enterprises. This broad base of support provides credibility, but the biggest impact of open-source infrastructure is the democratization of big data capabilities, especially when coupled with cloud computing. On-demand cloud services have obviated the need for many organizations to maintain dedicated hardware infrastructure, and today, analyses on terabytes of data can be conducted at modest costs without the need for major capital investments in servers. These tools are now within the reach of many social scientists, transforming the types of analyses they are able to conduct.

Data science and machine learning

The modus operandi of many consumer Internet companies is to begin with a successful product and attempt to induce the following virtuous cycle: by observing user behavior, data scientists gain insight into how the product can be refined

and improved; this hopefully leads to a more engaged and expanded user base, which in turn yields more behavioral data to analyze, thus completing the cycle.

The ultimate goal of data science teams within such organizations is to promote certain types of user behaviors that are aligned with business objectives. For example, online retailers want to maximize the number of shoppers and revenue. Once users arrive at the site, this can be accomplished by making products easier to find, reducing shopping cart abandonment (e.g., users who add items to their shopping carts but never "check out"), offering product recommendations, and so on. Social networks wish to increase engagement and grow the user base. This can be accomplished by recommending content items (e.g., posts, stories, tweets, friends, etc.) that the user may find interesting, for example.

The success of these activities can often be objectively quantified by metrics. For example, clickthrough rate is one such measure that captures the ratio between clicks (user actions) and impressions (opportunities for the user to take that action). We can measure the clickthrough rate of links on a landing page, the fraction of recommendations users click on, the frequency at which a product feature is invoked, and so on. For social networks, the number of active users is an important metric, as are session duration and total time spent on the site. For online retailers, total revenue, revenue per user, and other related metrics are obviously important.

What is the point of gathering such metrics? As Sir William Thomson (better known as Lord Kelvin) declares, "To measure is to know." The adage is usually followed up by its corollary, "If you cannot measure it, you cannot improve it." Once a metric is defined, it is possible to objectively determine which one of many competing methods is superior. For example, we can compare a number of content recommendation algorithms using clickthrough rate. We can see which alternative interface leads to longer user sessions, which phrasing of a welcome message generates more user sign-ups, or which type of mobile alert more effectively compels users to log in and check for updates. Of course, coming up with the right metric is often a challenge, and poorly defined metrics can lead to perverse incentives that negatively impact an organization's success; a simple example is a metric that cannibalizes long-term growth for short-term gains (Kohavi et al. 2012). Nevertheless, appropriate metrics provide clear definitions of success and failure.

Metrics allow organizations to compare alternatives, and in the online context, these comparisons are usually conducted via a process called A/B testing (Kohavi, Henne, and Sommerfield 2007; Kohavi et al. 2009). Although there are many nuances in properly executing such controlled experiments (Kohavi et al. 2012), the overall idea is fairly simple. Users are randomly assigned to one of two variants (sometimes called "buckets"): the control, which is typically the existing version of a particular feature, and the treatment, which is typically the new feature or intervention being evaluated. Metrics (per above) are then gathered and statistical tests are conducted on the collected data to determine if there is a statistically significant difference between the two conditions. If so, it can be asserted that the treatment is better than the control condition. In this manner, data scientists can compare the effectiveness of different recommendation

algorithms, page layouts, checkout flows, banner messages, and so on. Many organizations have A/B testing frameworks that allow data scientists to "plug in" and conduct such experiments in a streamlined fashion (Tang et al. 2010): this represents critical software infrastructure in today's competitive environment, as the speed at which a company can improve its online offerings is often limited by its ability to iterate through successive A/B testing cycles.

As previously mentioned, predictive analytics is often used to describe the activities of operationalizing insights into data products. That is, first we try to understand how users behave; then we introduce interventions that attempt to influence their actions in a manner that is consonant with the metrics; A/B testing then closes the loop to tell us if we have been successful. Machine learning has become the tool of choice for building such interventions. Although the complexities of machine learning are myriad and beyond the scope of this article, different techniques share in the idea of capturing statistical regularities in some observable input (called features or "signals") and some output (what social scientists would call the dependent variable), for the goal of making predictions on unseen data. Thus, machine learning is fundamentally about generalizing the past, via a process called "training," to (hopefully) predict the future. As a simple example, a spam classifier is trained on a corpus of spam emails based on features of those messages, for example, their textual content, originating IP addresses, and so on. Particularly useful features are referred to as "strong signals," which indicate (relatively) high correlations with the outcome we are trying to predict. A spam classifier induces statistical regularities from these examples, which are captured in a model and used to make predictions on emails it has never seen before.

Machine learning techniques are ubiquitous today in online environments. They are responsible for keeping our inboxes free from spam, personalizing the content of websites that we visit to better cater to our interests, verifying that a credit card purchase is legitimate, suggesting that we reconnect with an old friend, and recommending that we consider a competing product when we shop online. Machine learning to a large extent owes its success to big data—every time a user clicks a link, an ad, or a recommendation, the interaction is recorded and added to the vast stores that are exploited to train future models. Researchers have discovered that, all things being equal, the effectiveness of a model increases with the amount of training data (Banko and Brill 2001; Brants et al. 2007; Halevy, Norvig, and Pereira 2009). Machine learning, therefore, contributes to the virtuous cycle of big data: better models lead to higher quality products and attract more users, which generate more training data that can be further leveraged to improve the models. Of course, these improvements do not continue forever and we eventually reach a point of diminishing returns. Nevertheless, data volume remains an important driver of model quality.

Better Science versus Better Engineering

The central thesis of this article is that much of the controversy about big data stems from a fundamental confusion between what is science—understanding

human behavior and offering explanations of social phenomena—and what is engineering—building more effective computational artifacts as measured by some well-defined metric. I believe that many criticisms of big data arise from confusion over this distinction. For this discussion, we might say that understanding, while potentially helpful, is not necessary for prediction. That is, big data engineering provides us with the tools to predict behaviors without necessarily understanding the underlying sociological phenomena.

To dive into more detail, let us recap some of the criticisms that have been leveled against big data:

- Correlation does not imply causation. Data analysis at scale can detect correlations between a multitude of signals, but it cannot tell us if the correlations are meaningful. Furthermore, if we look multiple times for correlations between variables, our statistical significance tests are expected to find bogus relationships based purely on chance.[1]
- Big data cannot replace the scientific method. They cannot provide a substitute for a well-formulated hypotheses and the training required to interpret the nuances of a particular data collection method.
- Signals gathered by big data techniques are often the output of instruments that have not been properly calibrated and verified as producing reliable data.

These statements are all absolutely true and are precepts of what we would consider "good science." Indeed, big data techniques can tell us about correlations but offer no help in untangling causation. Big data are no substitute for hypothesis-driven scientific discovery, and many of the types of data gathered are dependent on the idiosyncrasies of the data collection system. However, I contend that these criticisms are largely irrelevant when the objective is to build an effective computational artifact. It may sound tautological, but if the goal is to improve a particular metric, the only thing that matters is improving that metric.

Let us explore in more detail this distinction between science and engineering with the following observations:

1. Global average temperatures are (negatively) correlated with the number of pirates.[2]
2. Per capita chocolate consumption correlates with the number of Nobel laureates (by country) (Messerli 2012).
3. Cloud cover correlates with stock market movements.[3]

Most people would dismiss the first correlation as not meaningful. No serious climate scientist would include "number of pirates" as an input to climate simulations, as the purpose of climate modeling is to better understand and explain the interactions among the atmosphere, oceans, land masses, and human activities. The predictions generated by the models are important only insofar as they lead to insights about climate phenomena. Accurate forecasting without satisfactory

explanations in terms of underlying physical mechanisms has little scientific value.

The second example comes from an article published in the prestigious *New England Journal of Medicine* (Messerli 2012). It was meant as a joke, but the author proposed a causal mechanism: "flavanols, which are widely present in cocoa . . . [seem] to be effective in slowing down or even reversing the reductions in cognitive performance that occur with aging" (p. 1562). Thus, "chocolate consumption could hypothetically improve cognitive function not only in individuals but also in whole populations" (p. 1562). Regardless of the original intent, this article has inspired at least one serious research study, which disproved the hypothesis (Maurage, Heeren, and Pesenti 2013). Despite the lack of scientific validity, from an engineering perspective, if one wanted to build a model that predicts intelligence, this may nevertheless be a feature to consider including (more below).

The final example comes from a recent retrospective event involving Peter Brown and Bob Mercer, two pioneers in data-driven approaches to natural language processing, particularly machine translation. Their seminal contributions happened at IBM in the late 1980s and early 1990s, but soon after, the researchers left IBM and helped to build one of the world's most successful hedge funds. In discussing signals relevant for prediction, Brown says, "It turns out that when it's cloudy in Paris, the French market is less likely to go up than when it's sunny in Paris. That's true in Milan, it's true in Tokyo, it's true in Sao Paulo, it's true in New York. It's just true." He isn't bothered by the seemingly inexplicable connection and continues, "We have . . . like 90 PhDs in Math and Physics, who just sit there looking for these signals all day long. We have 10,000 processors in there that are constantly grinding away looking for signals." It is clear from his description that understanding the underlying causal mechanisms is not a priority. He nicely summarizes: "It's ruthless. Either your models work better than the other guy's, and you make money, or they don't, and you go broke."[4]

This final example starkly illustrates the contrast I am drawing between science and engineering. The types of analyses that data scientists engage in are much closer to hedge funds trying to make money than climate scientists trying to understand physical phenomena. For the purposes of optimizing a particular metric, standard techniques in machine learning such as cross-validation, feature selection, and regularization are remarkably powerful in determining which signals are useful and which are not. Thus, it is better to let the machine learning algorithm "do its job," since it is far more sensitive to statistical patterns than human intuition. In fact, the standard approach in machine learning is to throw in "the kitchen sink" of features (and the cross product of features with other features, plus the features transformed or discretized in some way, and so on) and let the model sort them out.

But of course, training a model is akin to predicting the past. This is where A/B testing comes in: if a relationship is nonexistent and the perceived signal is actually noise, prospective (i.e., future) predictions will fail. Similar to Feynman's declaration that "nature cannot be fooled," real-world user behavior is the ultimate validation. That is, of course, assuming proper A/B testing methodology

(see Kohavi et al. [2009, 2012] on the nuances of properly executing such controlled experiments). Why does cloud cover impact stock prices? Who knows; but the more important point is, who cares? If the feature "works"—in the strict sense that it makes accurate future predictions (i.e., makes money for the hedge fund)—we should use it without prejudice, even if it appears inexplicable.[5]

I am not denying the usefulness of explanatory models but simply saying that they may not be necessary if the goal is to build a computational artifact that accomplishes a well-defined task. In short, prediction does not require understanding, although understanding certainly may help. Indeed, the most effective machine learning approaches arise from features that encode plausible causal linkages—such approaches combine the strength of humans (domain knowledge) and machines (extracting statistical regularities). As a hypothetical example, suppose that a data scientist notices a correlation between global average temperatures and online sales volume. She could simply throw that feature into the model, as with the cloud cover example, and move on to another task. The machine learning algorithm might be able to extract a small signal that leads to a miniscule increase in effectiveness. Or alternatively, the data scientist might dig a little deeper: perhaps an explanation might be that online sales tend to increase when it is cold because shoppers prefer to stay home. From this insight, follow-up analyses might show that this effect is most pronounced for North American and European cities, and only during the winter months in the northern hemisphere. This would lead to much stronger signals that could be integrated into the model, and the discovery about connections between weather and consumer behavior could be generalized and applied to other scenarios, such as weather-based strategies for email campaigns. This is an example where the desire to make accurate predictions is complemented by an attempt to understand the phenomena involved, which leads to better solutions.

Nevertheless, the inability to make sense of a signal should not prevent us from introducing it into a model, letting the machine learning algorithm "figure it out," and allowing A/B tests to determine the ultimate success. While such an approach may not be intellectually satisfying, the history of human knowledge is full of examples where the ability to accomplish certain tasks predates our understanding of the underlying principles that govern it. For example, steam engines and powered flight came before our understanding of thermodynamics and aerodynamics, respectively. Cathedrals soared into the sky long before structural engineers formally understood static and dynamic loads. Mendel was able to "predict" the color of peas in the next generation without knowledge of the molecular underpinnings. The pragmatist says that as long as we can build something useful today, understanding can come tomorrow.

In contrast to "building a better mousetrap" and a focus on engineering artifacts that achieve higher effectiveness on well-defined metrics, many social scientists are using big data to understand the complexities of human behavior, such as how individuals form and maintain social ties and the dynamics of influence and power. This, of course, is a fundamentally different endeavor; since we desire understanding, improving metrics is neither necessary nor sufficient to make a contribution to knowledge. As an illustrative example, I contrast two studies that

are concerned with link formation on Twitter, but have completely different goals. Romero and Kleinberg (2010) studied the directed closure process in hybrid social-information networks. Their work is grounded in the well-known process of triadic closure in social networks (Rapoport 1953; Granovetter 1973), which they extended to directed information networks (since social networks are typically modeled as undirected graphs). They hypothesized a mechanism for this process based on information "copying" and related their analyses to preferential attachment (Albert and Barabási 2002). Contrast this work with a description of Twitter's Who-to-Follow (WTF) service (Gupta et al. 2013)—the production system that provides recommendations to users and is responsible for creating millions of connections daily. Note that although both studies are concerned with edge formation on the Twitter follow graph and use similar metrics for evaluation, they have vastly different aims. Romero and Kleinberg use "closure rate" to validate their model of link formation, and thus advance our understanding of the directed closure process. In the WTF service, clickthrough rate is simply a metric used in A/B testing to determine if one algorithm is better than another. Some of the features used in the WTF algorithm are inspired by constructs that sociologists would recognize; for example, one could argue that the random walk algorithms used by WTF operationalize notions of closure. However, there are many features that are not grounded on any sociological principle and are more akin to the "cloud cover" feature discussed above. The WTF service attempts to balance the science and engineering aspects, which is productive since there is a vast literature on tie formation from the social sciences. However, there is no confusion about the ultimate goal: to increase the density of the Twitter follow graph (to ultimately promote more engagement among users). In contrast, the work of Romero and Kleinberg is about science. Once this distinction becomes clear, the criteria for assessing quality are straightforward: In the former case, does the model reveal some insight about user behavior in online networks? In the latter case, have we improved the relevant metric?

Let us return to the criticism from Lazer et al. (2014), who charged that "big data hubris is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis" and that "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data." I believe that this is an unfair attack on a straw man, as no experienced data scientist would seriously advocate throwing out everything we know about a subject and using only big data for studies where a deeper understanding of the phenomenon is the goal.[6] Well-trained data scientists know to be careful in confusing correlation with causation. They know that explanation is possible only via the formulation of testable hypotheses that propose a causal link from observations to outcomes. Vast data mining and other bottom-up efforts to see "what the data say" are guides to hypothesis generation and should not be mistaken for explanatory models. The unfair criticism aside, the quote above gets it right: big data are a supplement, not replacement, to traditional techniques.

I believe that much of the criticism of big data stems from confusion over the science versus engineering dichotomy. To further highlight this, consider a

parallel from efforts to apply computational techniques to language. It is clear that there are two very different endeavors: one is to build computational systems that process natural language to accomplish a task, for example, speech recognition or machine translation. The other is to take advantage of computational models to better understand the human capacity for language, for example, simulations of language change in speaker populations or models that try to explain phoneme perception. Perhaps because these different studies are carried out by largely disjointed communities, there is very little confusion about the objectives. Consider Google's approach to improving machine translation by building statistical language models on terabytes of web data (Brants et al. 2007). Few would criticize the technique by saying that no human could possibly read that much text, and so the approach is not a cognitively valid model of how humans perform translation. While the above statement is factually correct, most people would recognize the argument as a non sequitur. How humans translate languages is irrelevant to building systems that accomplish the same task.

Let us apply the same logic back in the domain of large-scale human behavior: in a recent editorial,[7] the well-known computational biologist Steven Salzberg attributed the failure of Google Flu (Ginsberg et al. 2009) (using Google search volume to predict flu outbreaks) to the fact that ordinary people do not really understand the complex virology of influenza: often when they search for flu-like symptoms, they do not really have the flu. While Salzberg is correct about the complexities of the illness, I think his criticism misses the point. The question is not about the public's understanding, but whether there is any signal in search queries that can be useful for early disease detection and public health. As such, the only important measurement of success is whether Google Flu makes correct predictions about the future (and the associated costs). In this respect, Lazer et al. (2014) point out several flaws (which is the right argument to be having), but they concede that "greater value can be obtained by combining GFT [Google Flu Trends] with other near-real-time health data. . . . For example, by combining GFT and lagged CDC [Centers for Disease Control and Prevention] data. . . . We can substantially improve on the performance of GFT or the CDC alone." This is exactly the point I am making. Of course we should take advantage of all features that are available, including data from traditional sources, but at the same time, we should not discount a signal (search query volume) simply because we do not completely understand it or because we have questions about its validity.

Conclusion

I conclude by considering another criticism from Lazer et al. (2014), which is that there is a lack of transparency associated with many of the signals that are derived from big data. If we analogize such signals as readings from scientific instruments, we should be concerned about the lack of calibration, evidence of reliability, and so on. For example, in the Google Flu study, there is no disclosure of exactly which search terms are considered in measuring the search volume;

furthermore, the Google search algorithm is constantly evolving and would likely affect the signal itself. Of course, Google can internally validate its measurements, so this criticism is really about who has access to the data, an issue also raised by boyd and Crawford (2011).

This "data divide" is by no means a new problem. Ever since the dawn of writing, when Sumerian scribes committed information to clay tablets that were then collected in the archives of the kings, society has had asymmetries between those who have access to information and those who do not. It is not even clear that big data have exacerbated the divide; consider, for example, the tiny fraction of the population in Europe that had access to books in the Middle Ages.

Although in an ideal world data would be freely accessible to researchers, this is simply not possible given the realities of today's competitive environment. For many organizations, the data form the "secret sauce" and thus are jealously guarded. Available APIs are burdened by usage restrictions, for example, rate-limiting by Twitter, which preclude easy bulk collection. However, academic researchers should give their industry colleagues more credit in understanding the importance of data sharing and collaborative efforts. Beyond genuine privacy concerns in many cases, hesitation is not borne from malice or apathy but simply a matter of competing priorities—most corporate cultures do not reward external engagement, which means that there is little incentive to help researchers with data requests. Nevertheless, as companies mature, they generally become more receptive to external engagement; for example, witness the introduction of Twitter's data grant program for researchers.[8]

In an attempt to be constructive, I offer two suggestions. First, collaborations between academics and industry partners provide access to valuable data. However, building meaningful collaborations requires substantial investments from both parties: I spent two years working at Twitter to build the relationships that support mutually beneficial collaborations. Scores of faculty spend sabbaticals at Google, Yahoo, and Facebook, which lead to joint research and publications, so getting access to data is possible. In the absence of direct faculty involvement, students who spend summer internships in industry provide bridges back to their home academic institutions, which often lead to lasting impact in the student's research (and today this is commonplace). The scarcity of talent in the technology sector means that industry is eager to recruit the best and brightest, and they understand that an internship is often a prelude to future permanent employment.

Second, even without direct collaborative ties to an organization, much meaningful work can still be accomplished. Let me elaborate from both the engineering and the science perspectives: one important aspect of engineering is the ability to build useful artifacts from unreliable components. In particular, much of software engineering is concerned with composing abstractions that are only accessible through well-defined interfaces. In a service-oriented architecture, which represents one common approach to building large systems today, the services are assumed to be unreliable. Why can big data signals not be treated the same way? If the signal is useful, it should be exploited, even as a black box. This

highlights the importance of continuous monitoring and testing, so that we can identify when a signal stops working and remove it from our models.

From a science perspective, the potential unreliability of big data signals forms a subject worthy of inquiry. In many scientific disciplines that involve use of complex instruments, studies of the instruments help to establish the context for their proper use. Indeed, Galileo made substantial improvements to the refracting telescope before turning his creation toward the skies. Why can the same ideas not be applied to big data "instruments"? If GFT lack transparency and replicability, why not search for alternative correlates that are more easily obtainable, say from Twitter (Paul and Dredze 2011)? If we find Twitter's API too restrictive to reconstruct data about a particular topic, can we devise a principled methodology that works within the constraints but is able to maximize coverage (Ruiz, Hristidis, and Ipeirotis 2014)? What are the effects of data sampling strategies on the ability to accurately characterize a phenomenon on social media (De Choudhury et al. 2010)? These and related studies are critical to helping the field better understand the limitations of big data and place themselves as contributions to knowledge.

I believe that most of the complaints about lack of data access are actually about lack of easy data access. Many researchers want to visit a website, download a dataset, and immediately begin analysis. In some disciplines this is indeed possible, for example, taking advantage of the University of California, Irvine's, machine learning repository.[9] Unfortunately, this is an unrealistic expectation when working with social media data. Data collection represents an integral part of any empirical research effort, especially ones that study human behavior, and I believe the discussion above outlines a productive agenda for overcoming obstacles related to data access today.

Recognizing the dichotomy between science and engineering is critical to properly situating big data research, particularly as it relates to the social sciences. Researchers should clearly articulate whether the goal of their work is to "understand the human condition" or to "build a better mousetrap," as this distinction lays out the criteria by which contributions should be assessed.

Notes

1. See http://xkcd.com/882/.
2. See http://www.venganza.org/about/open-letter/.
3. See http://cs.jhu.edu/~post/bitext/.
4. See Brown and Mercer (2013).
5. An important related point that is often neglected is that a model needs to be continuously validated over time to ensure that it keeps making accurate predictions. If we cannot explain why a particular signal is helpful, we are also unlikely to know when that signal stops working.
6. It may perhaps be the case that reports on the successes of big data are often too flippant without careful consideration of the nuances of a complex problem, and media reports are partially to blame in their oversimplification. However, this is a question of tone, not substance.
7. See http://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/.
8. See https://blog.twitter.com/2014/twitter-datagrants-selections.
9. See http://archive.ics.uci.edu/ml/.

References

Albert, Réka, and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74:47–97.

Backstrom, Lars, Paolo Boldi, Marco Rosa, and Johan Ugander. 2012. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference (WebSci '12)*, 33–42, Evanston, IL.

Banko, Michele, and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, 26–33, Toulouse, France.

Becla, Jacek, and Daniel L. Wang. 2005. Lessons learned from managing a petabyte. In *Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR 2005)*, Asilomar, CA.

boyd, danah, and Kate Crawford. 2011. Six provocations for big data. Paper presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society," 21 September 2011.

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 858–67, Prague, Czech Republic.

Brown, Peter, and Bob Mercer. 2013. Oh, yes, everything's right on schedule, Fred. Transcription of a discussion at the EMNLP 2013 Workshop on Twenty Years of Bitext. Available from http://cs.jhu.edu/~post/bitext/.

Dean, Jeffrey, and Sanjay Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th USENIX Symposium on Operating System Design and Implementation (OSDI 2004)*, 137–50, San Francisco, CA.

De Choudhury, Munmun, Yu-Ru Lin, Hari Sundaram, K. Selçuk Candan, Lexing Xie, and Aisling Kelliher. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, 10–17, Washington, DC.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457:1012–14.

Granovetter, Mark S. 1973. The strength of weak ties. *American Journal of Sociology* 78 (6): 1360–80.

Gupta, Pankaj, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. WTF: The Who to Follow service at Twitter. In *Proceedings of the 22nd International World Wide Web Conference (WWW 2013)*, 505–14, Rio de Janeiro, Brazil.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24 (2): 8–12.

Hey, Tony, Stewart Tansley, and Kristin Tolle. 2009. *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Kohavi, Ron, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2012)*, Beijing, China.

Kohavi, Ron, Randal M. Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: Listen to your customers not to the HiPPO. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2007)*, 959–67, San Jose, CA.

Kohavi, Ron, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery* 19 (1): 140–81.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: Traps in big data analysis. *Science* 343 (6176): 1203–1205.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science* 323 (5915): 721–23.

Maurage, Pierre, Alexandre Heeren, and Mauro Pesenti. 2013. Does chocolate consumption really boost Nobel award chances? The peril of over-interpreting correlations in health studies. *Journal of Nutrition* 143 (6): 931–33.

Messerli, Franz H. 2012. Chocolate consumption, cognitive function, and Nobel laureate. *New England Journal of Medicine* 367 (16): 1562–64.

Paul, Michael J., and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 265–72, Barcelona, Spain.

Rapoport, Anatol. 1953. Spread of information through a population with sociostructural bias: I. Assumption of transitivity. *Bulletin of Mathematical Biophysics* 15 (4): 523–33.

Romero, Daniel M., and Jon Kleinberg. 2010. The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, 138–45, Washington, DC.

Ruiz, Eduardo, Vagelis Hristidis, and Panos Ipeirotis. 2014. Efficient filtering on hidden document streams. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*, Ann Arbor, MI.

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2010)*, 17–26, Washington, DC.

Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The anatomy of the Facebook social graph. arXiv:1111.4503v1. Available from http://arxiv.org/pdf/1111.4503.pdf.

Zachary, Wayne W. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33 (4): 452–73.