I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

The Two Cultures and Big Data Research

EYTAN ADAR

Abstract: The ongoing struggle in the integration of Big Data methodology into the social science "toolkit" is due, in no small part, to the gulf between the "two cultures" of research. Those that produce and work with explanatory models (the first culture) criticize the primarily predictive models (the second) produced as part of Big Data research. While often the debate does not acknowledge the role of models as a fundamental point of contention, it nonetheless underlies much of the discourse. By better appreciating this difference and finding ways in which to integrate models, Big Data social science will become better integrated in the general social science research practice. The goal of this work is to critically examine the existing gap, the consequences of its existence, and possible resolutions.

I. INTRODUCTION

The use—current and potential—of Big Data for social science research has brought to light a significant number of tensions between different academic sub-communities. The lack of an agreed-upon definition for Big Data has exacerbated the situation and created a major expectation gap. Both critics and proponents of Big Data research are responsible for highlighting anecdotal evidence of the failure or success of Big Data—on occasion citing the same case study (e.g., Google's "flu trends work" (Ginsberg et al. 2008)). At the extremes, Big Data in the view of some researchers is either (a) entirely new, —the ”[B]ig [D]ata will revolutionize research" community (Anderson 2008, 2) or (b) nothing new—the ”[B]ig [D]ata is just data" community (Few 2014, 1; Nafus and Sherman 2013, 1786-

1787).[1] Neither extreme, nor the practice of identifying anecdotal evidence to argue for one or the other, is a particularly helpful practice as it fails to address the underlying differences and their roots in the debate over scientific practice. The worst offenders of this "extremist" viewpoint are often media who are tasked with translating scientific discourse (and who often benefit from the appearance of a "battle"). However, the press often models itself on the existing culture clash—one being fought in public, rather than academic, contexts (Marcus and Davis 2014; Hidalgo 2014, 1).

Specifically, this clash emerges due to different modes of scientific practice and objectives. Traditionally, the social science community has focused on *explanatory models* (and *constructed data*),[2] whereas those in the computational sciences have targeted *predictive models* (and *observational* or *"found"* data). The latter often represents "Big Data research." The perspective that these two modes of inquiry are somehow incompatible is detrimental to scientific progress and blinds academics to the benefits of other perspectives.

The current discourse is harmful to the way in which academia, the public, and policy makers engage with, and adopt, Big Data practices. Furthermore, the confounding of academic and industrial practice of Big Data under one name further fuels the debate. It makes it difficult to hold the position of being for academic Big Data practices, but against aspects of corporate practice (which amplify ethical concerns) since, from a high-level, they appear to be one and the same. This particular feature will hopefully become irrelevant as we move away from the "Big Data" moniker and adopt more specific ways of discriminating between techniques, applications, and values.

Regardless, my contention—and the focus of this paper—is that the differences between the two modes of research are important to understand and, more critically, the ability to utilize both is critical to our understanding of social phenomena in both the theoretical and applied sense. While Big Data veers toward the predictive/observational due to various structural reasons, it nonetheless offers significant benefits to those working in domains that demand theoretically grounded, explanatory models. Rather than fixating on these extremes, there is an opportunity to identify ways in which both modes of work can be used effectively in complementary ways. The traditional practice of social science—at least the parts for

[1] As with any debate, there are those that take a more nuanced stance that both considers the benefits and highlights the concerns (Boyd and Crawford 2012, 671). Nonetheless, the framing of many of these concerns, I believe, relates to the models of inquiry.

[2] Although the broad range of social science sub-disciplines make it difficult to completely generalize, this is nonetheless a widely held belief.

which there is a Big Data analog—will need to find some way to incorporate the large-scale computational techniques of Big Data.

II. EXPLANATORY AND PREDICTIVE

In some sense, the distinction between explanatory and predictive models is highly nuanced. The difference is deeply philosophical in its relation to the objective of scientific inquiry and an active area of debate (Breiman 2001, 199; Shmueli 2010, 289). The similarity between the two is most simply represented by Figure 1, which abstractly captures both approaches.

Figure 1: The abstract model (adapted from Breiman 2001)

The specific elements in both model types are the same, but the objective of the inquiry is often different. Broadly, we seek to understand the relationship between some variable X and some other variable Y. A predictive model focuses on predicting Y given X (see Figure 2) i.e., identifying a function f(X)→Y. The function may be some type of regression (e.g. linear or logistic) or classifier. Notably, the function need not emulate the true underlying relationship between the two variables. In fact, the relationship need not be causal at all—"nature" may be driving both X and Y simultaneously. A correlational relationship, where X is a leading indicator for Y, may be as desirable as a true causal model in the case of predictive modeling. For example, one could predict the probability of purchasing some item (Y) given observed tweeting behavior (X).

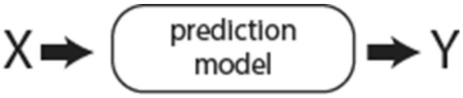

Figure 2: Prediction (adapted from Breiman 2001)

A distinguishing feature of predictive models in the context of Big Data is that data is often *found* (or *observed*) rather than *constructed*. In contrast, an explanatory model (Figure 3) is focused on the causal process (the "nature") that mediates the input and output. Most often, through a careful collection of data from controlled experiments, self-report (e.g., surveys), or other methodologies, it becomes possible to isolate the relationships between the different scientific objects (often *theoretical constructs* in the social sciences). These theoretical constructs and associated theories are key features of this approach. One could, for example, model purchasing behavior by understanding the influence of factors such as brand recognition, reputation, or social influence. These operationalized objects can be isolated and tested in a way that lends confidence—often through specific regression models—to a specific causal interpretation that connects to an underlying theory. Though often these causal models are not, in fact, true causal models in the statistical sense, the underlying theory is viewed as justification of the causal argument (Shmueli 2010, 290). Experimentally derived evidence is often held in highest regard, but it is not always possible to achieve, and the specific method of deriving data is not as critical as the controlled construction of the data.

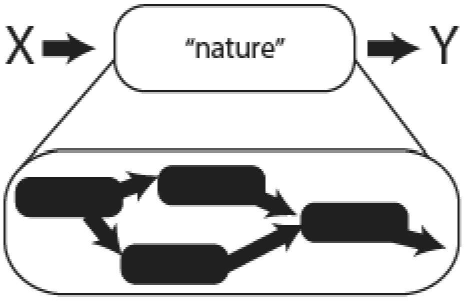

Figure 3: Explanation (adapted from Breiman 2001)

A predictive model *may* seek to identify key leading indicators that can be leveraged to predict some outcome (other tasks may involve classification or repairing missing data). While we would often like these predictors to be theoretically motivated, they can be obtained through other mechanisms (e.g., data mining). In some sense, a predictive model is about simply being able to predict something, not about determining *why* something happens. Conversely, a good explanatory model may not work well as a predictive model. While the causal relations would, in theory, imply that Y can be predicted through X as the "flow" between them is captured, the confidence in that prediction may be incredibly small. Furthermore, there are many situations where it is difficult or impossible to practically capture X. For example, X and Y may occur too closely in a temporal sense or X may be hard to operationalize (e.g., requiring a complicated psychological assessment survey).

Despite their similar structures, a good predictive model need not be a good explanatory model. This subtlety is by no means obvious, and from the perspective of the press, the public, some policy makers, and even some scientists, the relationship can be completely misunderstood.

A. *Prediction and Forecasting*

A further source of confusion derives from the use of the word "predictive." As predictive, in conventional language, implies some assertion about the *future*, this has created an ambiguity that extends from academia to the public. Predictive models are often "predictive" in the statistical case, where a function can map some input X to some output Y. For example, a regression fit to some observed dataset may be such a function. However, Y need not be a future observation. We may accurately model the relationship between tweets and votes during a *particular election cycle* (e.g., for every *x* units of increase in tweets for a particular candidate, we observe a change of *y* in the number of votes). This model may simply capture the relationship in an interesting way or allows us to infer missing data. However, for various reasons, this *particular* model may not work for a *future* election. To distinguish between the two types of predictions, I will adopt the notion of *forecasting* as a specific sub-type of predictive modeling, one that may work ex ante.

This difference in fact forms part of the arguments against Big Data. Critics of Big Data techniques often point to the inability of predictive models to function as forecasting tools. For example, the use of Twitter as a sensor for votes and the accuracy of Google Flu Trends have been heavily criticized (Metaxas, Mustafaraj, and Gayo-Avello 2011, 165; Marcus and Davis 2014, 2-3). The models work well at a particular time or for a particular election cycle, but fail beyond this. The failure to forecast is perceived as a general failure of predictive models. Perhaps worse, those criticizing the models often focus on singular failures as signs of a more general problem.

On the other hand, those creating the models often oversell the forecasting power of the models (Huberty 2012, 1). Or, put more generously, they are unable to counteract the misinterpretations of what they meant by "predictive." While ideally predictive models are "stable" across time, and therefore function to forecast, the reality is that they often require active recalibration (Lazer et al. 2014, 1204). Thus, while the "signs" in the models may remain correct (i.e., more tweets or more searches are correlated positively with more votes or more illness), the magnitude of those relationships may dramatically change over time.

Rather than resorting to ambiguous definitions to further a point, it is far more mutually advantageous to provide crisper definitions of the goals of the models, their nature, and their limitations (as well as recognizing the existence of "fixes" to those limitations).

B. *The Black Box and the Unknown*

To summarize our simple model (Figures 1-3), explanatory models seek to understand the "black box," whereas predictive models are content leaving it as "unknown." The distinction exists more broadly depending on the domain. In statistics, for example, this has taken the form of the "two cultures" debate (*data models* versus *predictive models*). This debate is roughly analogous to the explanatory/predictive split in the social sciences, if not in precise structure than certainly at a high level (Breiman 2001, 199). Other fields (e.g., epidemiology) have more successfully bridged the gap by using the different methods as part of an overall toolbox. Similarly, those in the fields of biology and chemistry, with the new focus on large-scale experimentation (e.g., microarrays), have been forced to build new ways of incorporating big data into their scientific practice. Interestingly, the different disciplines have adopted different strategies to incorporate the different methods. The social sciences have been much slower to create this bridge.

The educational system for social science academics has advocated one view to the exclusion of the other. The result has been a deeper philosophical divide that has contributed to the strong reactions between the communities and has materialized in the extreme opinions around Big Data. This is not to claim that the early adopters of the predictive/observational side were completely prepared for the nature of Big Data, which nonetheless requires new methods, but they were certainly more prepared. In particular, the "realities" of the data gathered through observational means forced the construction of new techniques and research designs. While the lack of "controls" in this type of data is deeply uncomfortable to the explanatory side, it fit naturally in the predictive side.

III. UNOBTRUSIVE AND NON-REACTIVITY IN BIG DATA

The choice of the modeling style is inexorably tied to the nature of data itself. The "data" part of Big Data—at least when it comes to the kind most suitable for social science research—is often collected not through experimental procedures, but rather through observation—the logging of behavioral traces (often in "uncontrolled" environments). Occasionally this collection is explicit, but more likely it is a secondary artifact of the information production and consumption process of people living their ordinary digital lives.

Social scientists have names for this kind of data—*unobtrusive* or *non-reactive* data. Though the names are often used interchangeably, the difference is somewhat critical in the case of Big Data social science as it forms another point of contention (Tufekci 2014, 1-10). Specifically, unobtrusive observational data most often refers to data collected without the awareness of the subject. Non-reactive data is data that is collected in a way that will not influence the subject's behavior.

Arguably, one would like both properties to be true—the subject is unaware and unaffected. Unfortunately, while data can be non-reactive at the time of the collection and simultaneously unobtrusive (e.g., collecting political tweets), the outcome of this research often provokes a reaction in the study population (Tufekci 2014, 1-3; Cueni and Frey 2014). In the political tweets example, the revelation that tweets are somehow correlated with election results may lead a political operative to produce additional tweets—an attempt to "game" the system. A more obvious example may be a fund trading on stocks mentioned through social media. Having found some leading indicator in the signal, the company acts on the signal but consequently "corrupts" the signal when they move the market. Any advantage of the signal may be reduced over time.

While "undesirable" from the perspective of science,[3] reactivity is often part of the reality of Big Data as often the data being collected is being collected specifically to predict and modify whatever behavior was being observed. For example, search logs are collected to improve search, consequently causing changes in search behaviors and influencing the models. Purchasing logs are used to model and modify purchasing behavior and so on. More subtly, changes in the underlying system that are beyond the control of the modeler can create havoc in the predictive power of the system. A criticism of the Google Flu Trends project raised this potential issue—that the modifications of search behavior and user interfaces by one team at Google would change the data observed by the data miners in significant and uncontrolled ways (Lazer. et al. 2014, 1204).

Social science research using nonreactive/unobtrusive information began in the 1960s (Webb et al. 1999, 5) and often involved creative collection of information that could be transformed into an instrument for the construct in question, but was often a "by-product" of something else. For example, to judge which painting is the most

[3] If data is reactive, the stability or generalizability of the model can be brought into question—an undesirable characteristic for both predictive and explanatory models.

popular in an art gallery, it was proposed that the rate in which tiles were replaced in front of different pieces was a good indicator. To find out how frightening a children's story was, one only needed to measure the tightness of the circle children made around a fire during a campfire telling. To identify on which radio station a car dealer should advertise, they could simply look at what their customers were tuned to when bringing their cars in for servicing. Modern examples of this have yielded more significant commercial and scientific outcomes. For example, to judge the likely revenues of a Wal-Mart store, it was proposed that the number of cars in the parking lot, as identified in a satellite image, was a good proxy for customers and, therefore, business (Javers 2010). The large scale "garbage project" looked through the physical trash left curbside by residents in order to understand patterns of consumption (Rathje and Murphy 2001, 13-14). This kind of data is noticeably "distant" from the construct of interest. Arguably, other measures would better tell us about the popularity of art pieces (cameras in front of every piece or exit interviews), the scariness of a story (galvanic skin response sensors and heart rate monitors on every child), or the best place to advertise or what is being consumed and why (survey instruments and focus groups). However, often the best measures are expensive or hard to create.

Despite their obvious utility in certain contexts, unobtrusive measures have received negative attention since they were first introduced (Webb and Weick 1979, 211-212). The data is viewed as noisy and, because it is somewhat uncontrolled, thought to be less scientific. This argument has extended to the Big Data era as many of the data products being used in this type of research have a similar structure. As such, they have been called everything from "found data" (a more neutral term) to "data exhaust" or "data fumes" (a more polarizing term).

It is worth noting that techniques for utilizing observational data in general, and unobtrusive data specifically, have long been extended and refined. Multiple models, methods, and instruments are often brought to bear on a study to test multiple angles. Noise is accounted for explicitly and, in fact, often this noise is embraced as another source of data (Webb et al. 1999, 39). Regardless, the tidal wave of observational Big Data does not appear to be dwindling. Simply ignoring it because it is more difficult to use in conventional ways is a limiting perspective, which we would do well to overcome.

A. *The Focus on Predictive*

It is worth briefly addressing the sources of observational data that is being made available for social science research which is often the data produced by large corporate and government entities. Many of these entities make this data directly available to researchers—though not always in easy to collect ways (e.g., rate limited APIs, data in PDFs, limited sharing rules, etc.). To focus the present discussion—and because they are a likely a source of a majority of Big Data datasets—we can narrow our analysis to Internet-based producers (such as Twitter, Facebook, Microsoft, Google, eBay, Reddit, and Kiva).

The most immediate objection to data obtained from these sources is that the lack of specific documentation and control over many elements of the data collection pipeline, as well as the particular biases of the corporation, will constrain the research questions that can be asked. Though some research has been attempted to reverse-engineer the properties of the data (Morstatter et al. 2013, 1), in many situations the particular stream is undocumented and can change at any time.

While there are many companies that are positioned to help social science research, there is rarely a perfect alignment between commercial and academic research interests. Clearly, the agenda of companies will focus the kinds of questions they ask and consequently the kinds of data they capture: can I understand my customers and predict what they will do next? For example, if they search for "apples" on my search engine, what will they click on next? Or, if they searched for apples, how likely is it that they will purchase an apple? And finally, given that they are about to buy an apple, what ads do I show them so they will buy a specific kind of apple? Many of these questions can be answered directly through *observational* data collection or log data (e.g., behavioral traces of end-user interactions). Given enough data, and some randomization, it is possible to identify a large enough cohort to achieve predictive aims. Put another way, we can observe enough people searching for "apple" to begin to develop a predictive model of how likely they are to buy an apple (many search vendors can track a user beyond the search site through cookies or so-called "toolbar" data to identify a "hit" on a shopping site).

While observational data is present in great quantities, often *experimental* data collection can also be undertaken—so called A/B testing. Under A/B testing, different service variants can be used to identify those that produce a "better" behavior. The standard reductionist example of this type of experiment is whether the color or text of the "buy now" button, or the advertisement on the site, encourages more active purchasing behavior. For a company, there is a tradeoff in using this kind of technique as it more directly answers specific questions (which color button is better?), but may not be broadly useable in "log mining" contexts. While established companies have learned to keep track of interface versions or which A/B condition their end-users are exposed to, this is not true for all and may impact the predictive task. This last point has a direct consequence to researchers using of logs in secondary contexts. Very rarely are researchers aware of the particular conditions under which data was collected and incorrect assumptions may result in questionable study validity.

Finally, it is worth pointing out that, even though A/B testing has the appearance of standard scientific experimental procedures, it would be a mistake to confuse the two. While some A/B testing can be used downstream for explanatory modeling, this is not always the case. For example, A/B tests are not necessarily theory-driven and a test condition can be designed that compares alternatives, but does not test a particular theory. Specifically, our hypothetical search vendor may show two different ads (half of the end-users get one, the other half get the other) to determine which one is clicked on more. There is no theory in this experiment—it is simply the testing of alternatives. This is not to say that an A/B test cannot become a "scientific" experiment, but rather that companies are not incentivized for this.

Facebook's recent "emotion manipulation" experiment reflects an interesting case study (for many reasons) in that it arguably fit into both categories (Kramer et al. 2014, 8788). The study manipulated the exposure of end-users to positive and negative posts (randomly hiding posts with particular keywords) to identify the influence on the end-user's own posts. The study both had a potential benefit to the company (e.g., understanding how your posting behavior varies from your "neighbor's" may be used to design interfaces or algorithms that encourage posting behavior), as well as scientific inquiry (e.g., how emotional contagion may work (Fowler and Christakis 2008)). This is, unfortunately (again, for many reasons), a rare instance of research where academic and corporate interests were aligned closely enough to create an interventional study (likely to become even rarer given the particularly negative response to the study).

It is possible, of course, for social science researchers to develop their own infrastructures for collecting Big Data. At the University of Michigan, for example, we have begun to develop and deploy the MTogether system,[4] an observational and interventional platform built into desktop and mobile platforms that tracks social media use and can "manipulate" a user's experience. The initial releases were designed to leverage the alumni and fan base for Michigan—big in the "little b" sense. However, platforms such as MTogether are hard to get right; they are expensive to develop and deploy, and the sustained uptake or growth of the platform is still uncertain. Successful companies are able to provide extrinsic and intrinsic motivations to their customers. It is an open question whether research platforms can also achieve the same results.

[4] University of Michigan, "MTogether", http://www.mtogether.us (2014).

To conclude, the particular incentives of corporations have notably biased Big Data toward the observational/predictive side of the spectrum. Social science researchers who are unwilling or unable to engage with this type of data are often left out of Big Data research. This by no means indicates that research using other types of data or methodologies are, or will become, irrelevant, but this does point to a particular fracture between sub-communities that needs to be mended or at the very least more deeply understood.

IV. THE VIRTUOUS CYCLE

Having focused on the differences in research philosophies that have led to divisiveness around Big Data research, it is worth considering mechanisms by which these may not only be bridged, but also linked in a mutually beneficial way.

A. *Education*

The introduction of Big Data techniques has led to the creation of new opportunities for students to gain experience. Programs such as the Interdisciplinary Committee on Organizational Studies (ICOS) at the University of Michigan offer a yearly "boot camp" on Big Data for researchers in the social sciences (largely PhD students). The week-long intensive seminar is co-sponsored by both the Colleges of Engineering, Literature, and Science & Arts. The goals of the camp, and others like it, is to train students to leverage Big Data resources (e.g., Twitter's data stream) and tools (e.g., iPython). The camp exists, in part, because "[i]t is a good bet that within a few years, a standard part of graduate training in the social sciences will include a hefty dose of 'how to make use of big data,' just as statistical analysis is a standard part of such training today." (ICOS 2014) Other disciplines have similarly incorporated the computational techniques that are necessary for Big Data research into their educational culture and there is no reason to believe that the social sciences will be unable to do so.

B. *Model Feedback Loops*

There is a great deal of potential for the inclusion of the predictive models that are a key part of the Big Data world. The relationship of predictive and explanatory models has long been a topic of discussion, though not always in the context of social science research or Big Data specifically. Shmueli (Shmueli 2010, 289) has reviewed a number of ways in which explanatory and predictive models can work in concert. We briefly summarize a few key points about the value of predictive modeling raised in this work (Shmueli 2010, 292):

New data contexts – Predictive models often serve as a way of engaging with new datasets about which no hypotheses have been formed. New media data (e.g., social media) does not always fall into the classical forms (what other forms of communication are limited to 140 characters?). Predictive models can help to understand these datasets.

1. *New measure generation* – Predictive models can be leveraged to discover new measures and test their operationalization.

2. *Connection to practice* – The use of predictive models can test the practical/applied validity of explanatory models. Furthermore, an explanatory model that is used in a predictive context, but which does not perform as well as a predictive model, may point to the potential to improve theory.

3. *Tests for competing theories* – Due to their assessment mechanisms (e.g., accuracy), predictive models often have a more natural support for comparing competing theories.

Conversely, theory and explanatory models can serve to improve predictive models as well. For example, theoretical models allow us to test predictive outcomes and assess the long-term potential of a predictive model. The relationship between the S&P 500 and the Bangladeshi butter supply (where the latter can predict the former) has long been viewed as an example of data mining gone awry (Leinweber 2007, 16). Though the butter supply has predictive power, it clearly does not make much rational sense and is likely to fail when applied as a forecasting tool. Theory about market systems would likely tell us that this particular indicator should be removed from consideration in the predictive model.

Additionally, explanatory models have the practical purpose of helping in the feature-engineering task that is often a key factor in the design of predictive models. Having some insight into the causal model can guide the creation of these features and can focus the effort of the researcher. A recent paper at the International Conference on Weblogs and Social Media (ICWSM) received a best paper award (Lietz et al. 2014, 7) for an analysis of the German elections using Twitter data. The work did not attempt to predict the election, but rather study it from the perspective of theoretical constructs that were carefully operationalized and tested on Twitter data. These "features" are hopefully more robust than the traditional "bag-of-words" approach that has been used in Twitter analysis and are likely to be more robust over time—leading to true forecasting, rather than simple prediction, *and* deeper insights about the social processes around elections.

C. *Triangulation*

No particular research methodology—interviews, surveys, lab experiments, or any other applied within the social sciences—is completely immune to criticism. Each has limitations with no obvious "fix" (e.g., bad self-reports on surveys, interviewer bias, ecological validity, biased sampling, small-N's, etc.).

The existence of multiple approaches should be viewed positively as it allows us, as scientists, to test and validate our ideas of how individuals and social systems work. In part, this validation through multiple studies is a better standard—one that should be attractive to those trying to understand social phenomena in a robust way. Researchers who take advantage of nonreactive measures (Webb et al. 1999, 16) have long been aware of the fact that multiples of everything (instruments, methods, models, etc.) leads to more robust findings. It is worth taking this advice to heart and finding better ways to tie techniques together.

V. ETHICAL CONCERNS

The integration of Big Data research into the social science toolbox has led to extensive discussion about ethical considerations. Though much of this criticism has been targeted more broadly at corporate interests, it is nonetheless critical to understand the context of research as well, especially as the two are often intertwined either in perception or in reality. In particular, as academic social science researchers begin to either leverage the output of corporate interests (e.g., Twitter) or collaborate with their corporate counterparts, ethical concerns become crucial to understand and address. The recent Facebook emotional contagion study vividly demonstrates this (Kramer et al. 2014, 8790). The extremely high variance in the response from the public, academics, the press, ethicists, and corporate interests demonstrates that we have not yet converged on a completely satisfactory solution that can balance the demands of scientists, the public, and corporate interests.

The design of policy and the education of the public are clearly crucial, but along with these it is clear that more discussion needs to occur within the scientific community. The opinion on ethicality of the Facebook study was roughly split, with computational scientists arguing that it was ethical and the more traditional side arguing that it was not. This is unfortunate for a number of reasons, one being that it reinforces the tension between these communities. Hopefully, the ongoing attempts at a productive dialog around Big Data ethics will provide a foundation for broader collaboration.

VI. CONCLUSIONS

The contemporary definition of Big Data around the three V's—volume, velocity, and variety (Gartner 2014)—is a beguiling one. It captures some of the high level properties of the data, but fails to engage with the nuances of how it is used and what scientific insights can be gleaned from it. Simplifying definitions and anecdotes collected from the press cannot be the best way to move forward for social science research. Rather, critiques of the specific methods and engagement with the limitations and benefits of different models and data sources would seem to be better choices. Our ability to do so ensures that we are able to educate future social science researchers and expand their toolbox of techniques. This is a matter not only of pursuing "the right tool for the job," but also of leveraging the complementary strengths of the different modes of scientific inquiry. This is not to say that we should stop with understanding the specific properties of the tools. The use of Big Data has re-opened debates within academia around ethical data use. It is clear that, in addition to developing new ways to mine data or integrate theory, these concerns should form a key part of the discussion.

A. *Acknowledgements*

I would like to thank Peter Shane for organizing the Big Data Futures conference and giving me the opportunity to think through these issues. I would also like to thank my colleagues Nicole Ellison, Sarita Yardi Schoenebeck, and Cliff Lampe who provided me with significant feedback on drafts of this work.

REFERENCES

Anderson, Chris. 2008. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired*, June 23.

boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15, no. 5: 662-679.

Breiman, Leo. 2001. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical Science* 16, no. 3: 199-231.

Cueni, R. and Frey, B.S. 2014. “Forecasts and Reactivity.” *CREMA Working Paper, Center for Research in Economics, Management, and the Arts* no. 10.

Few, Stephen. 2014. “Big Mouths on Big Data.” *Visual Business Intelligence: A blog by Stephen Few*, April 30, (3:30 p.m.), http://www.perceptualedge.com/blog/?p=1891.

Fowler, J. H., and N.A., Christakis. 2008. “Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study.” *BMJ* 337.

Gartner. 2014. “Big Data, IT Glossary.” accessed July 31, http://www.gartner.com/it-glossary/big-data/.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2008. “Detecting Influenza Epidemics Using Search Engine Query Data.” *Nature* 457, no. 7232: 1012-1014.

Hidalgo, Caesar. 2014. “Saving Big Data from Big Mouths.” *Scientific American*, April 29. 2014.

Huberty, Mark Edward. 2013. "Multi-cycle forecasting of congressional elections with social media." Proceedings of the 2nd workshop on Politics, elections and data. ACM.

ICOS. 2014. “ICOS Big Data Summer Camp, University of Michigan.” accessed July 31, 2014, http://ibug-um.github.io/2014-summer-camp/.

Javers, Eamon. 2014. “New Big Brother: Market-Moving Satellite Images.” *CNBC*, August 16.

Kramer, Adam D. I., Jaime E. Guillroy, and Jeffrey T. Hancock. 2014.“Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks.” Proceedings of the National Academy of Science 111, no. 24: 8788-8790.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 342, no. 6176: 1203-1205.

Leinweber, D. J. 2007. “Stupid Data Miner Tricks: Overfitting the S&P 500.” *The Journal of Investing* 16, no. 1: 15-22.

Lietz, H., C. Wagner, A. Bleier, and M. Strohmaier. 2014. “When Politicians Talk: Assessing Online Conversational Practices of Political Parties on Twitter.” Proceedings of International AAAI Conference on Weblogs and Social Media, AAAI.

Marcus, Gary and Ernest Davis. 2014. “Eight (No, Nine!) Problems With Big Data.” *The New York Times*, April 6.

Metaxas, Panagiotis Takis, Eni Mustafaraj, and Daniel Gayo-Avello. 2011. "How (not) to Predict Elections." *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom).*

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2003. “Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose.” Proceedings of International AAAI Conference on Weblogs and Social Media, AAAI.

Nafus, Dawn and Jamie Sherman. 2013. "This One Does Not Go Up To Eleven: The Quantified Self Movement as an Alternative Big Data Practice." *International Journal of Communication* 8: 1784-1794.

Rathje, W. L. and C. Murphy. 2001. *Rubbish!: the Archaeology of Garbage*. University of Arizona Press.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25: 289-310.

Tufekci, Z. 2014a. “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological

Pitfalls. Proceedings of International AAAI Conference on Weblogs and Social Media, AAAI.

Tufekci, Z. 2014b. "Facebook and Engineering the Public," The Message, June 29, 2014, accessed July 31. https://medium.com/message/engineering-the-public-289c91390225.

Webb, E. J., D. T. Campbell, R. D. Schwartz, and L. Sechrest. 1999. *Unobtrusive Measures*. vol. 2. New York: Sage.

Webb, Eugene and Karl E. Weick. 1979. "Unobtrusive Measures in Organizational Theory: A Reminder." *Administrative Science Quarterly* 24, (1979).