SYMPOSIUM

# Drawing Inferences and Testing Theories with Big Data

**Jonathan Nagler,** *New York University*
**Joshua A. Tucker,** *New York University*

This paper considers two primary issues about big data: (1) big data as a sample, and (2) a particular type of big data— that is, social media data. We argue that if we can learn things from little data, then we certainly can learn from big data. We think that big data offers a tremendous opportunity that little data does not. We argue that having more data is an opportunity for, not a constraint on, testing theories of political behavior.

We address three aspects of big data as a sample. Regardless of how *big* big data is, it is a sample drawn from a population. First, this means that we have to consider how representative the sample is of the population that we claim to be studying.[1] Second, we have to consider whether the population from which the sample was drawn is a population of interest and/or whether the population it was drawn from generalizes to other populations. Third, because big data is big, we have to adopt more caution than typically is the case when interpreting our measures of statistical precision. Big data drives *sampling error* toward zero but does not reduce other errors associated with inferences drawn from a sample. That big data drives sampling error toward zero makes big data different from little data in that we must be careful not to be fooled by our own *estimated* precision.

Although there are many types of big data available, we consider what we can learn from social media data. Big data comes in many forms, including individual-level campaign contributions (Bonica 2013), large corpora of text from traditional print media, records of legislative debate, and judicial opinions. Big data also comes from new activities, such as counts of online searches. We think that social media data is particularly valuable to political scientists for at least three reasons. First, it allows us to observe a rich set of data on those networks in which people choose to embed themselves. Second, social media data allows us to observe the comments that people freely make about politics. Thus, if we are interested in knowing what people think, we do not need to ask—they volunteer this information on social media. Third, because social media data also is big data, it offers advantages over traditional types of political science survey data. Because so many people are on social media, we can obtain large samples from subpopulations of interest, whether they are defined by geographic characteristics (e.g., residents of a particular state), demographic characteristics (e.g., young black men), or political characteristics (e.g., Democrats). Fourth, because people use social media to communicate about a variety of political activities, including protest activities and political mobilization, we can use it to learn about these strategies more generally. Thus, social media data comprises a source of big data by providing information about peoples' opinions regarding politics, their choices of information sources, the nature of their associations, and their strategic choices in certain political situations. As political scientists, we think we can learn something from this.

## BIG DATA AS A SAMPLE

First, we want to make a few simple points. In some ways, big data is not different than other data. If we can learn things from little data, we ought to be able to learn things from big data. So we still can learn useful things from a 2x2 table using big data. And we still can learn useful things *and make causal inferences* from a well-specified model testing clearly defined hypotheses using big data. We even can run experiments with big data: randomizing treatment across respondents and observing outcomes. We also can make bad mistakes with big data, just as we can with little data.

One of the things that has happened with big data is that computer scientists have discovered big datasets related to politics. This is both a good thing and a bad thing. As social scientists, we are interested in stating and testing hypotheses. We typically test our hypotheses by trying to draw statistical inferences. We find a sample, analyze data, and draw a statistical inference about the parameters of the population from which the sample was drawn. So far, the computer scientists who have analyzed big data about politics have not worried so much about this. They are very good at worrying about whether their estimates are applicable *to their sample:* they are much more thorough than the typical political scientists at out-of-sample testing. However, in this context, "out of sample" means split the sample into many parts, and determine whether analyzing one part of the sample can be applied to the other parts of the sample. But of course, since the sample is split at random, this is a means of establishing how well we can forecast, *assuming our sample is representative of the population.* However, this is not a way to check whether the sample is representative of the population.

However, as political scientists, we typically are concerned about the properties of the sample. Specifically, is it representative of the population? This is a major problem in traditional survey data—a problem that has waxed and waned over time. We have come a long way from the *Literary Digest* days. Standardized techniques have been developed for drawing random samples. Although we have struggled with the banal



doi:10.1017/S1049096514001796

problem of who is more likely to answer the telephone, until relatively recently, telephone surveys seemed to be well understood. Cell phones have made surveys more difficult to conduct, and response rates have plummeted across both types of telephones. Furthermore, Internet surveys have ushered in the world of bigger (but not big) data for survey research. Serious survey firms (e.g., Knowledge Networks and Polimetrix) have pioneered methods to survey people over the Internet, resulting in representative samples of populations of interest (e.g., registered voters, or the adult population). Of course, the Internet also has introduced the age of the convenience sample, which is anything but representative.

With big data comes the illusion of big precision. In one often-repeated paradigm, scholars estimate a model of vote choice for a given election using the National Election Study and they publish a result with a t-statistic more than 2. We then conclude that we are 95% certain that our result would hold (to the extent that "would hold" means the sign would not change) if we had data on the entire set of voters for that election. Through sleight of hand or carelessness, we can pretend that our result would hold for other elections and that we have found a universal truth because our result is "statistically significant." However, we have not found a universal truth; rather, we have found something that was probably true in the one election that we observed. Does this result generalize to other elections? We have no idea. We must take it on faith that the election we observed was randomly drawn from all possible elections. Or, we must believe that the model we specified applied to all elections. Perhaps it does, or perhaps we omitted a variable that changes from election to election.

For example, we could estimate a model of voter turnout in the 2008 election using the 90,000 or so observations of the Current Population Survey. Based on our estimates, would we be certain that older people, *ceteris paribus*, are more likely to vote than younger people? Of course not. We would be very certain of this only for the 2008 election. We could state this because we have a representative sample of all potential voters in that election. But perhaps there was something specific to the election that caused older people to vote more than younger people? If we want to know if this were generally true, we would need to observe many more elections.[2] Or, we would have to assume—and assuming is not knowing—that our correctly specified model accurately describes the data-generating process in every election, not only the one on which we estimated it.

With big data, if we estimate a quantity and compute a standard error, it will appear as if we are very certain of something. This is the beauty of dividing by $\sqrt{N}$. However, all that we are very certain of is that if we had even more data drawn from the same population as our sample, and we estimated the same quantity, we would obtain the same result. None of this suggests that the quantity we estimated was meaningful. And this is the danger of inference with big data. The error that we observe most easily is our *sampling error*. However, big data drives the *sampling error* toward zero while having no effect—relative to small data—on other causes of total survey error. Total survey error would include (1) error induced by the failure of the sampling design to generate a sample that was truly randomly drawn from the population, (2) failure to handle unit nonresponse, and (3) measurement error induced by any survey instrument.[3]

*None of this suggests that the quantity we estimated was meaningful. And this is the danger of inference with big data. The error that we observe most easily is our* sampling error.

Luckily, the discipline (or parts of it) has recognized the fallacy of asterisks and the worshipping of statistical significance just in time for the deluge of big data and big t-statistics. Part of the credit for this goes to the Bayesians, but part of it is simply common sense. Therefore, if we report standard errors computed from big data, while we are still computing a value associated only with sampling error, at least we may be less likely to conclude that our findings are validated by strings of asterisks indicating statistical significance at traditional levels.

However, big data itself comes in many flavors, and sometimes it can appear that we know the population. We may know the entire number of Google searches on Google for a particular term or set of terms. But presumably, that is only a proxy for some facts of political interest. If we consider Google searches to indicate interest in something, then if we see the number of searches for "impeach Obama" increases, for example, we might think that interest in impeaching the president is increasing. It might be increasing among the people likely to use Google to search for political facts. If 95% of the voting-age population is likely to use Google to search for political facts in this way, and if the number of searches for "impeach Obama" doubles, then using the method of bounds, we can be quite certain that interest in impeaching the president has risen among the voting-age population, *if* we believe that those searches express interest. However, most questions in big data are not so clear cut. This simple example suggests something to which we return in the following discussion: Why not survey people to determine whether those who searched for "impeach Obama" are more likely to want to impeach the president than those who did not search for "impeach Obama?"

## SOCIAL MEDIA DATA

We now turn to social media data, with a particular focus on Twitter data. By using the term "social media data," we are referring to data that has distinct characteristics. First, it is

posted online by the mass public. Second, people can freely choose the set of people from whom they receive information; that is, they choose the information network in which to embed themselves. Third, people can comment on information that others post. Fourth, there may or may not be anonymity. We consider Twitter data. A user on Twitter can follow people who they know in the offline world; there is no anonymity. However, John Smith-Public on Twitter also can follow Jane Smart-Political-Analyst because many of his friends follow her. However, John SP cannot be certain if Jane SPA is a 45-year-old woman in Los Angeles—as her profile claims—or a 17-year-old in China pretending to be her. Furthermore, tweets by Mary Good-Environment might actually be from a coal-industry lobbyist trying to make moderate-sounding comments to suggest that global warming is not a problem. In either case, John SP knows only that the tweets sent over time by Jane SPA and Mary GE represent those sent by the same person (or entity).

Twitter data is big data: worldwide, there are more than 270 million active Twitter accounts.[4] Everyone on Twitter has chosen whom to follow, thereby publicly revealing their network to us. This network includes people that they follow who can reveal something about their politics as well as the people who they follow for presumably nonpolitical reasons (e.g., entertainers, sports figures, friends, and relatives) but who nevertheless may provide them with political information. We learn who people are following because this information is publicly available on Twitter; we do not have to depend on their recall or their description of the four or eight people with whom they communicate most often. This means that we can observe which types of political information people are choosing to obtain. If we want to study the effect of political information on attitudes and/or behavior, this is valuable data. If we only want to examine which information is likely to be believed, this also is valuable because we can observe what information people choose to pass on via retweets.

We believe that it is fruitful to think of Twitter as unstructured survey data. In most surveys used by political scientists to study political behavior, an analyst asks the questions using a structured format, which most often requires closed-ended responses. We can think of Twitter as having completely open-ended responses, in which the analyst has not even posed a question. Thus, it is an unfiltered look into what people choose to say about politics. We can learn which issues that people on Twitter think are important over time, and we can do this in at least two different ways. As analysts with substantial expertise in the area of study, we can search for terms on Twitter that we think correspond to topics of interest. For example, we can search for the term "immigration" to determine how often it is used. This could serve as a proxy for how much the set of Twitter users care about the political issue of immigration. However, this suggests obvious potential pitfalls. The term "immigration" could be used in contexts other than political contexts. Thus, faith in counting, without human verification from a small subsample of tweets with the term "immigration" to confirm that a majority is about politics, could result in substantial errors. Of course, depending on the data, confirming that "a majority" of such tweets is about politics might not be adequate. Imagine that we observe a 10-percentage-point increase in the number of tweets per week containing "immigration." If we know that only 70% of the tweets concerned the political topic of immigration, then it would be risky to conclude that a 10-percentage-point increase in tweets containing the term "immigration" signaled an increase in concern for the political topic of interest. The increase could represent an increase in nonpolitical uses of the term "immigration."[5]

However, in addition to searching for occurrences of terms preselected by an analyst, we can search in the text for terms that appear together in our attempt to identify topics. This is common in the machine-learning community via Latent Dirichlet Analysis (LDA): a corpus of documents is created (here, the set of tweets) and analyzed based on an underlying probabilistic model to assign words to topics and, thus, tweets (based on word content) to topics. The analyst then looks at the words in each "topic" and decides whether the "topic" revealed by LDA in fact represents a substantive political topic.

Thus if, as political scientists, we want to know what people on Twitter are talking about, and what political events or stories or topics they are talking about, we have ways to do so. This brings us to the obvious question: do we care what people on Twitter are discussing? If we can say that they are representative of some population we care about (e.g., voting-age adults), then we would care. We also care if there are simply enough of them to care about, but it would make inference much harder. For instance, if we measure the opinion of the top 20% of US households based on income, and we observe that it changes on an economic issue, then we could relate that to politics: we would know that the opinion *of the rich* is changing.

*One potential solution to the problem of making opinion on Twitter politically relevant is to look at how opinion varies across Twitter users based on politically relevant characteristics of the users.*

But, Twitter users are not a politically relevant category; thus, simply knowing that opinion on Twitter is changing (or even what it is in a static sense) is not as useful for us as political scientists.[6] Thus to make use of opinion measured on Twitter, we want to be able to somehow relate Twitter users to the population of interest.

One potential solution to the problem of making opinion on Twitter politically relevant is to look at how opinion varies across Twitter users based on politically relevant characteristics of the users. Very little information about Twitter users is

provided by default. But, we can estimate the gender of Twitter users based on their name with a high degree of accuracy. And Barberá (2014) showed that we also can accurately estimate the ideology of many Twitter users. Only about 5% of tweets are geo-coded, which allows a very accurate estimation of location. However, based on self-reported place, we can accurately predict the state a user resides in for about 70% of users. This means that, for example, we can identify that among Twitter users, immigration is talked about more in Arizona than in Idaho. And we also can identify that among Twitter users, immigration is talked about more by people on the political right than on the left. While there is no guarantee that those relationships hold in the voting-age population, it is not a bad hypothesis that they do.

Another way that opinion on Twitter can be politically relevant is to observe how it changes in response to political events; again, this can be refined by observing changes across different groups of users. Perhaps Twitter's greatest strength as a data source is that it is unprompted by survey questions, which means that it provides an unfiltered measure of whatever people are choosing to discuss. A long-standing question in the study of representation has been: Who leads and who follows? Do elected representatives set the agenda and constituents follow, or do elected representatives respond to issues about which their constituents care? In research with our colleagues in the Social Media and Political Participation laboratory at New York University on this question, we examined tweets by politically interested Twitter users (i.e., those who followed political figures) and tweets by members of Congress during a 14-month period (Barberá et al. 2013). This allowed us to compare topics that the mass public was talking about with those that members of Congress were discussing and to determine which group talked about a topic first (i.e., led) and which group talked about it after the other group (i.e., followed).

We want to point out some inferences not to draw from Twitter data. It is easy to look at Twitter data and notice a high correlation between the proportion of tweets for congressional candidates and the proportion of votes for them. First, inferring any causality from this would obviously be a bad idea. The winning congressional candidates spent more money than their opponents, had more supporters in their district, had higher name recognition, and so forth. Therefore, it is no surprise that a higher proportion of tweets were about them. We note that one could attempt to identify challengers who might be doing better than expected by unusually high numbers of posts on Twitter. However, this would require a model of the expected baseline number of tweets, which is likely to vary with district demographics as well as the demographic source of support for a candidate.[7]

As one of the specific purposes Twitter is used for is to disseminate information during protests, we also can learn about the strategy of protest by studying Twitter data. We can determine with whom protesters are trying to communicate by their choice of language (Tucker et al. 2014), and we can learn which people in protest networks are most likely to be influential in spreading information (Metzger et al. 2014).

**Twitter as a New Variable**

We point out that Twitter data is big data with two potentially interesting characteristics for political scientists. First, as we have previously discussed, Twitter data is a potential source of information about the political world, just as telephone and Internet surveys are. And it is very interesting data: repeated observations on millions of individuals collected over time. But second, Twitter is a potential variable itself in the political world. Twitter can provide much politically relevant information to people that has the potential to alter their political behavior. Twitter can provide people information about opportunities to participate politically, about the cost of participation, and about the potential outcomes of participation. Twitter can provide people information about opportunities for participation by informing people about everything from protest activities (i.e., where and when), to when a candidate forum or rally for a candidate is scheduled, to where a polling place is located. Without a means to know where a protest is, one cannot participate. Without a means to know where a candidate forum or candidate rally is, one cannot participate. Twitter is of course not the only means to know about these things. Thus, we need to be careful about drawing inferences about its causal impact, a point we return to below.

Twitter can also can provide people information about the cost of participation. If people are considering participation in a protest, they can find out whether or not the protest is well attended; or sparsely attended if they are worried that being one of the few people at a protest will make them targets of a hostile regime. They also can find out if a protest has the potential for violence: finding out that people are being shot at, or that large numbers of heavily armed military personnell have been seen heading towards the site of a protest, would give people valuable information about the likely costs of attending.

*As Twitter allows people to publicly share information about their views on politics, Twitter can also provide information to political elites about public opinion.*

Furthermore, Twitter can inform people about potential benefits of participation in the form of outcomes. If people learn on Twitter that one candidate advocates a policy of which they are in favor, there is more reason to participate and they are more likely to do so (Leighley and Nagler 2014).

Moreover, Twitter can provide not only information about political events and political elites but also what users' friends think about politics. This may lead to a world in which people

have an easier time making correct political decisions in their own self-interest. A long-standing paradox of democracy and political discourse is that citizens are presumed to vote for candidates closest to their political views. However, given the low likelihood of being pivotal in any election, it is difficult to explain why citizens would make the effort required to learn the candidates' views. The "cheapest" available solution is generally thought to be simply finding out for which candidate our politically informed friends are voting, with the implicit assumption that our friends have similar interests. Because Twitter provides a mechanism for people to broadcast their political views to their friends and for people to receive the political views of their politically interested friends, Twitter can be a way to achieve this information shortcut.

As Twitter allows people to very publicly share information about their views on politics, Twitter can also provide information to political elites about public opinion.[8] To find out if Twitter use is having an impact on political behavior and politics, we would want to do the obvious: compare Twitter users to non-Twitter users. We want to match a Twitter user to someone who is identical on all observable characteristics, and see if the Twitter user is likely to learn more during an election campaign (or during any period) than the non-Twitter user. We also can test persuasion: is the Twitter user more likely to change their intended vote choice (or choice of whether or not to vote) during the campaign than the non-Twitter user? This requires panel data: using traditional survey techniques to collect a set of respondents on Twitter and a set of respondents off Twitter, and comparing their responses over time. There is simply no way to draw causal inferences on the effect of Twitter by studying only those people on Twitter. However, we can indirectly observe people who were not on Twitter previously by comparing people on Twitter who joined at different times. In this case, causal inferences are possible, but more limited.[9] We think it is too early to claim that Twitter is having a large effect on political behavior; but we think it is important to study it in a rigorous manner to answer the question of whether it is having an effect, and, if so, what the effect of Twitter is and when it is likely to have the most impact.

## CONCLUSION

If we believe the old adage that "to err is human, to really mess things up requires a computer," then it follows that we could make big mistakes with big data. However, we also can learn much from big data; and following the same careful attention to the requirements of causal inference that we would follow with any observational dataset, we can indeed draw causal inferences.

## ACKNOWLEDGMENTS

This paper was prepared for presentation at the August 2014 APSA Annual Meeting in Washington, DC. The authors are members of the New York University Social Media and Political Participation (SMaPP) laboratory. The writing of this article was supported by the INSPIRE program of the National Science Foundation (Award #SES-1248077) and Dean Thomas Carew and the Research Investment Fund (RIF) of New York University. ■

## NOTES

1. Even if we have "all of the data," it is a sample of all of the data that could exist over time or all of the data to which we want to generalize. Therefore, conceptually, it is still a sample.
2. We have observed many more elections and find this to be true (Leighley and Nagler 2014).
3. See Biemer and Lyber 2010.
4. *Source*: Twitter's second-quarter 2014 results.
5. This is less likely a potential problem when we attempt causal inference and look at temporal relationships among more than one variable. In those cases, the typical assumption that the "noise" is expected to be orthogonal to the quantity of interest makes more sense.
6. We might think that the population of Twitter users discussing politics is politically relevant because it is this population that is disseminating political information via this new source.
7. On using Twitter as a source of big data to conduct polling, see Beauchamp (2013).
8. Of course, this assumes that political elites employ sophisticated data scientists to analyze the Twitter data.
9. Another way that Twitter can have an impact is when people on Twitter share information they receive with people off Twitter. See Vacarri et al. (2013).

## REFERENCES

Barberá, Pablo. 2014. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis*. DOI 10.1093/pan/mpu011.

Barberá, Pablo, Richard Bonneau, John T. Jost, Jonathan Nagler, and Joshua Tucker. 2013. "Is There Anybody out There? The Effects of Legislators' Communication with their Constituents." Paper presented at the 2013 Text as Data Workshop. London: London School of Economics and Political Science.

Beauchamp, Nick. 2013. "Predicting and Interpolating State-Level Polling Using Twitter Textual Data." Paper presented at the 2013 APSA Annual Meeting, Chicago, IL.

Biemer, Paul, and Lars Lyber (eds.). 2010. "Special Issue on Total Survey Error." *Public Opinion Quarterly* 74 (5): 817–79.

Bonica, Adam. 2013. "Ideology and Interests in the Political Marketplace." *American Journal of Political Science* 57 (2): 294–311.

Leighley, Jan E., and Jonathan Nagler. 2014. *Who Votes Now? Demographics, Issues, Inequality, and Turnout in the United States*. Princeton, NJ: Princeton University Press.

Metzger, Megan, Duncan Penfold-Brown, Pablo Barberá, Richard Bonneau, John Jost, Jonathan Nagler, and Joshua Tucker. 2014. "Dynamics of Influence in Online Protest Networks: Evidence from the 2013 Turkish Protests." Paper presented at the 2014 APSA Annual Meeting, Washington, DC.

Tucker, Joshua A., Megan Metzger, Duncan Penfold-Brown, Richard Bonneau, John Jost, and Jonathan Nagler. 2014. "Social Media and the #Euromaidan Protests." Paper presented at the Workshop on Mass Protest. London: London School of Economics.

Vaccari, Christian, Augusto Valeriani, Pablo Barberá, Richard Bonneau, John T. Jost, Jonathan Nagler, and Joshua Tucker. 2013. "Social Media and Political Communication: A Survey of Twitter Users during the 2013 Italian General Election." *Rivista Italiana di Scienza Politica* 43 (3): 381–410.