Drawing Inferences and Testing Theories with Big Data

Jonathan Nagler* **Josh Tucker**
Richard Bonneau **John Jost**

Abstract

We argue that having more data is an opportunity, not a constraint, on testing theories of political behavior. And we point out that much of big data is not just more data: but a new type of data. In particular, social media data provides us with repeated observations over time on many individuals; and such observations are the unfiltered opinions of those individuals, unbiased by any prompting by the analyst.

This is a DRAFT!!! Do not cite this version.
August 1, 2014

Paper prepared for presentation at the Annual Meeting of the American Political Science Association, August, 2014, Washington DC. The authors are all 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. * Corresponding author: jonathan.nagler@nyu.edu.

We want to make a few simple points. Big data is not different than other data in some ways. If we can learn things from little data, we ought to be able to learn things from Big Data. So we can still learn useful things from a 2x2 table using Big Data. And we can still learn useful things *and make causal inferences* from a well specified model testing clearly defined hypotheses using Big Data. We can even run experiments with Big Data: randomizing treatment across respondents, and observing outcomes. And we can also make bad mistakes with Big Data, just as we can with little data. In much of this paper we'll focus on a particular type of Big Data: social media data; within the set of social media data, we'll focus on twitter data. Twitter data is valuable because we observe what people say, and who they bother to repeat. It would be nice if we could connect this discussion to their opinions about the real world, and to try to draw inferences about the opinion of people *not on Twitter*, and it would be nice if we could see if being on twitter matters. We describe how to do that. Because if we can't do that, we might as well observe people playing Dungeons and Dragons (after all, lots of people do that too, but it does not have much impact on politics).

With Big Data comes the illusion of big precision. In one oft-repeated paradigm, scholars estimate a model of vote choice for a given election using the National Election Study and publish a result with a t-statistic over 2. We then conclude 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 slight of hand, or sloppiness, 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." But we have not found a universal truth: we have found something that was probably true in the one election we observed. Does it generalize to other elections? We have no idea. We would have to take it on faith that the election we observed was randomly drawn from all possible elections. OR, we would have to believe that the model we specified applied to all elections. Maybe it does? Or maybe we omitted some variable that changes from election to election.

An example of the above. I 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 my estimates, would I be very sure that older people, ceteris paribus, are more likely to vote than younger people? Of course not. I would only be very sure that *in the 2008 election* older people were more likely to vote than younger people. I could do this because I have a representative sample of all potential voters in that election. But perhaps there was something specific to the election that caused old people to vote more than young people? If I wanted to know if this were generally true, I would need to observe a lot more elections.[1] Or, I would have to assume – and assuming is not knowing – that my correctly specified model accurately describes the data generating process in every election: not just the one I estimated it on.

Luckily the discipline, or parts of it, have recognized the fallacy of asterisks and worshipping of statistical significance just in time for the onrush of Big Data and big t-statistics. Part of the credit for this goes to the Bayesians. But part of it is just common sense. So with Big Data, if we estimate something and compute a standard error, it will look like we are very sure of something. This is the beauty of dividing by $\sqrt{N}$. But what we are very sure of is that if we had even more data drawn from the same population as our sample, and estimated the same quantity, we would get the same result. None of this suggests that the quantity we estimated was meaningful.

One of the things that has happened with Big Data is that computer scientists have discovered it. This is a good thing, and a bad thing. As social scientists we are interested in stating and testing hypotheses. We usually 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 that the sample was drawn from. So far the computer scientists that have analyzed Big Data about politics have not worried so much about this. They are very good at worrying about their estimates being applicable *to their sample*: they are much

[1] We have observed a lot more elections, and it is true (**?**).

more thorough than the typical political scientist at out-of-sample testing. But here 'out of sample' means: split the sample into many parts, and see if 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 could forecast *assuming our sample is representative of the population*. This is not a way to check if the sample is representative of the population. And so far, computer scientsits generally have *not* worried if the sample they have is representative of a population of interest.

But as political scientists, we usually worry about the properties of the sample. In particular, is it representative of the population? This is a huge 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 were developed for drawing random samples. And while we still had to struggle with the banal problems of who is more likely to answer the phone, till relatively recently - telephone surveys seemed to be well understood. Cell phones have made life much harder. And internet surveys have ushered in the world of bigger (but not Big) data for survey research. Serious survey firms such as Knowledge Networks and Polimetrix have pioneered methods to survey people over the internet, and come up with representative samples of populations of interest (i.e., registered voters, or the adult population). The internet has also of course ushered in the age of the convenience sample that is anything but representative. People can recruit subjects

But with Big Data, we have data so big people often seem to forget that it is still (generally) a sample of something. But Big Data itself comes in many flavors, and sometimes it can appear we know the population. We might know the entire number of 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 take searches on Google to indicate interest in something, then if we see the number of searches for "impeach obama" increase, we might think that interest in impeaching Obama is rising. It might be rising, among the people likely to use

Google to search for political facts. Now if 95% of the voting age population is likely to use Google to search for political facts, and if the number of searches for "impeach Obama" doubles, then using the method of bounds we can be quite sure that interest in impeaching Obama has risen among the voting age population, *if* we believe that those searches express interest. But most things in Big Data are not so clear cut. And this simple example suggests something we return to below: why not survey people, and see if the people who searched for "impeach Obama" are more likely to want to impeach Obama than people who did not search for "impeach Obama"? But most things in Big Data are not so clear cut.

We now turn to social media data, with a focus on Twitter data in particular. By social media data we refer to data that has distinct characteristics. First, it is posted online by the mass public. Second, people can freely choose the set of people they receive information from. Third, people can comment on information that others post. Fourth, there may or may not be anoyminity. Consider Twitter data. A user on Twitter can follow people that they know in the offline world. There is no anoyminity there. But a John Smith-Public on Twitter can also follow Jane Smart-Political-Analyst because many of John Smith-Public's friends follow Jane SPA. But in fact John Smith-Public can not really be sure if Jane SPA is a 45 year old woman in Los Angeles, as Jane SPA's profile claims - or a 17 year old in China pretending to be a 45 year old woman in Los Angeles. Or, Mary Good-Environment might really be tweets by a coal industry lobbyist trying to make moderate sounding comments suggesting that global warming is not a problem. In either of these cases, all John Smith-Public knows is that the tweets sent over time by Jane SPA or Mary GE represent tweets sent by the same person (or entity).

So what can we learn with twitter data? First, it is Big Data. There are over 270 million active Twitter accounts worldwide.[2]

We point out that Twitter data is Big Data with 2 potentially interesting characteristics

[2]Source: Twitter second quarter 2014 results.

for political scientists. First, Twitter data is a potential source of information about the political world - just as phone surveys or 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 (where and when), to when a candidate forum or rally for a candidate is, to where a polling place is. Without a means to know where a protest is, one can not 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 will need to be careful about drawing inferences of its causal impact, a point we return to below.

Twitter can also provide people about the cost of participation. If people are considering attending a protest, they can find out if 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 can also 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 personell have been seen heading towards the site of a protest, would give people valuable information about likely costs of attending.

Twitter can also provide people about potential benefits of participation in the forms of outcomes. If someone learns on Twitter that one candidate advocates a policy the person likes, there is more reason to participate, and the person is more likely to do so (Leighley and Nagler 2013).

Twitter can also provide people with not just information about political events and

political elites, but it can inform them about what their friends think about politics. This could lead to a world where people have a much 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, but given the low likelihood of being pivotal in any election, it is hard to explain why citizens would put forth the effort required to learn the views of the candidates. The 'cheapest' available solution is to generally thought to be to simply find out which candidate one's politically informed friends are voting for, with the implicit assumption that one's friends will have similar interests. Since 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 means to achieve this information short-cut.

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.[3]

Now, moving away from the question as to what the mass public and elites can learn from Twitter, and thus how Twitter can influence political behavior – what, if anything, can political scientists learn from Twitter as a big source of Big Data? We believe that it is fruiful to think of Twitter as very unstructured survey data. In most surveys used by political scientists studying political behavior, the analyst asks the questions in a very structured format, most often with closed-ended responses. We can think of Twitter as being completely open-ended responses, where the analyst has not even posed a question. Thus it is an unfiltered look into what people choose to say about politics. Thus one thing we can learn from Twitter is what issues people on Twitter think are important over time. We can do this in at least two different ways. We, as analysts with substantial expertise in the area of study, can search for terms on twitter that we think would correspond to topics of interest. Thus we can simply search on the term "immigration" to see how often it is used. This

[3]This of course assumes that political elites hire sophisticated data scientists to analyze the Twitter data.

could serve as a proxy for how much the set of twitter users care about the political issue of immigration. This obviously suggests some potential pitfalls. The term "immigration" could be used in contexts other than political contexts. Thus faith in counting, without some human verification from a small subsample of tweets with the term "immigration" to confirm that a majority of them are about politics could result in substantial errors. And of course depending on the data, confirming that 'a majority' of such tweets are 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 only knew that 70% of our tweets containing "immigration" were about the political topic of immigration, then we would be living dangerously concluding that a 10 percentage-point increase in tweets containing the term "immigration" signaled an increase in concern for the political topic of interest. They might represent an increase in non-political uses of the term "immigration."[4]

However, along with simply looking for occurences of terms the analyst pre-selects, we can simply search in the text for terms that appear together in an attempt to identify topics. This is what is commonly done in the machine-learning community via Latent Dirichlet Analysis: a corpus of documents is created (here the set of tweets) and analyzed based on an underlying probabilistic topic model to assign words to topics, and thus tweets (based on word content) to topics. It is then up to the analyst to look at the words in each 'topic', and decide if the 'topic' revealed by LDA does in fact represent a substantive political topic.

Thus if as political scientists we want to see 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 talking about? If we can say that they are representative of some population we care about (say voting age adults), then we would care. We might also care if there are simply enough of

[4]This is less of a potential problem when we are trying to do causal inference and are looking at temporal relationships between more than one variable. In those cases, the usual assumption that the 'noise' can be expected to be orthogonal to the quantity of interest makes more sense.

them to care about - but it would make inference *much* harder. For instance, if we measure opinion of the top 20% of US households based on income, and observe that it changes on an economic issue - we could relate that to politics. We would know that the opinion *of the rich* is changing. But twitter users are not a politically relevant category, thus simply to know that opinion on Twitter is changing (or even what it is in a static sense) is not very helpful for us as political scientists. Thus to make use of opinion measured on Twitter, we are going to have to be able to relate Twitter users to the population of interest somehow.

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) shows that we can also estimate ideology of many Twitter users very accurately. Only about 5% of Tweets are geo-coded permitting very accurate estimation of location. But based on self-reported place, we can accurately predict the state a user resides in for about 70% of users. This means that we can identify that *among Twitter users*, immigration is talked about more in Arizona than Idaho. And we can identify that *among Twitter users*, immigration is talked about more by people on the 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 opinion on Twitter can be politically relevant is to observe how it changes *in response to political events*, and again this can be refined by observing how it changes across different groups of Twitter users. Perhaps Twitters greatest strength as a data source is that it is unprompted by survey questions. That means it provides an unfiltered measure to whatever people are choosing to discuss. A longstanding question in the study of representation is: who leads, and who follows? Do elected representatives set the agenda, and constituents follow; or do elected representatives respond to the issues that their con-

stituents care about? In research of ours on this question using Twitter we examined Tweets by politically interested Twitter users (those who followed political figures) and Tweets by members of Congress over a 14 month period. This allowed us to compare the topics the mass public was talking about with the topics members of Congress were talking about – and see which group began talking about a topic first (led), and which group started talking about a topic after the other group (followed).

We want to point out some inferences *not to draw* from Twitter data. It is easy to look at Twitter data and notice that there is a high correlation between the proportion of tweets for a congressional candidate, and the proportion of votes for a congressional candidate. 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 the district, had higher name recognition, etc.. 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. But to do so would require some model of the expected baseline number of tweets - likely to vary with the demographics of the district, as well as the demographic source of support for the candidate.

We now turn to the second aspect of Twitter as Big Data: it is not just Big Data, but Twitter itself can have political impact. To find out if Twitter is having an impact: we would need to do the obvious: compare Twitter users to non-Twitter users. We want to match a Twitter users 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 can also 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 compare responses over time. There is simply no way to draw inferences on *the effect*

of Twitter by only studying 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. Inferences here are possible, but more limited.

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. But we can also learn things with Big Data, and following the same sorts of careful attention to the requirements of causal inference that we would follow with any observational data set, we can draw causal inferences.