# A Case Study In Big Data And The Replication Crisis

**Kalev Leetaru** Contributor
I write about the broad intersection of data and society.

Shutterstock

The ever-growing world of "big data" research has confronted the academic community with unprecedented challenges around replication, validity and big data ethics in a world in which nearly every imaginable kind of data is readily available. From the "replication crisis" to whether traditional understandings of ethics should apply at all, innovation is frequently moving faster than questions about what kinds of research should be permitted and how to verify and validate the headline-grabbing findings pouring out each day. Yet, much of the conversation around things like replication or ethical review tends to focus on abstractions, hypothetical scenarios and policies. What does it look like to actually take a big data study published today in 2017 and attempt to replicate it and what might we learn about whether we should just give up on the notion of replication entirely?

Over the last several decades, federally funded academic research in the United States that involves human subjects has been governed by the concept of "institutional review boards" (IRBs), committees at each university that review all proposed research to ensure it complies with federal law and accepted norms and standards regarding the ethics and impact of that research on the persons being studied. One of the driving forces behind the creation of these ethical review boards were some of the high profile medical experiments conducted over the preceding decades on unwitting or unwilling subjects. Over time these IRBs grew to encompass more than just medical research, establishing purview over any academic research that involves the study of

human beings, including the social sciences, ranging from active psychological manipulation to passive mining of the vast trails of sensitive data residue modern humans leave that document the most intimate aspects of their lives.

It is the latter – the data mining of commercially available and open source public datasets that has generated considerable conversation with respect to how to effectively apply it in a data drenched world and to disciplines unaccustomed to thinking about human subject considerations.

Much of our daily lives today are lived on privately owned grounds in both the physical and digital worlds. As we walk down the street, dozens of private surveillance cameras watch us, while our cell phone and any of the dozens of apps we've installed or devices we wear map our precise location, monitor our heartrate and calorie burn, observe the text messages we send and the people we call, even the individuals we stop to chat with on the street. When we purchase groceries or fill a prescription, drive or take a cab, surf the web or post a photo of our children, all of this information is being streamed to private companies who have wide latitude and often Orwellian privacy policies permitting them to resell any of that data to advertisers and researchers at will.

With these riches of data, how do we decide what questions we should be ethically asking of all of this data and how can other scholars attempt to replicate the studies that are being published, given the scales and often proprietary data involved?



Crystallizing the questions, what does it actually look like to try and replicate a modern "big data" research study to validate its results?

A recent study on the ideological correlation between the social networks of journalists and the articles they write caught my eye after being covered in several media outlets and given the intense current interest in assessing the various kinds of potential bias confronting the news industry today. The paper involved mining half a million news articles authored by 1,000 journalists from 25 different news outlets, together with an analysis of their Twitter activity – a classic example of a modern "big data" social science study – and used this information to find "a modest correlation between the ideologies of who a journalist follows on Twitter and the content he or she produces."

After reading through the paper several times and following its citations, I had several methodological questions that didn't appear to be answered in the paper (academic papers typically have hard length limits that restrict the amount of technical detail they can include), so I wrote the authors asking if they could share their replication datasets for the paper and answer a few questions I had, since I had been unable to locate any replication or supplemental information on their website. Specifically, given that the results of their paper largely hinged on the specific set of 1,000 journalists selected for analysis, the comprehensiveness of the list of 500,000 articles, the makeup of

the 12,000 "highly politically active Twitter accounts" and the final set of ideologically weighted terms used to score the documents (given that the authors mention using domain expertise to manually adjust the final set of terms), I asked if they could share the list of URLs and Twitter handles (but not the content itself, since that would be a violation of copyright) and the list of keywords. I also asked if they could answer a few basic clarifying technical questions such as whether they filtered out quoted phrases (so that they were only looking at a journalist's own words) and which algorithm they used to extract each article from its surrounding web page (since an imprecise algorithm might scoop up advertisements or headers/footers that might include ideologically charged language unrelated to the journalist's own words).

While waiting for the authors to respond with the requested information and datasets, I began preparing to conduct the replication, involving tracking down the various software packages that I would need and rereading the references, formulas and workflows in detail so that I could precisely match what the authors had done.

In the process of performing the standard due diligence that precedes all replication studies (such as ensuring that all proposed activities would comply with relevant legal agreements, licenses and terms of use), I became aware that Section VII.A.4 (User Protection) of the Twitter Developer Agreement included the language "Twitter Content, and information derived from Twitter Content, may not be used by, or knowingly displayed, distributed, or otherwise made available to … profile individuals based on … political affiliation or beliefs…" In short, Twitter appeared to explicitly prohibit the very research that the paper was based upon.

Researching this further, it appears the company added this clause to its public Developer Agreement sometime on May 18, 2017, though it appears that its EULA may have included the cause as early as February of this year and a Salesforce license document from January 12, 2017 includes the prohibition, but applies it only to public sector customers, while an end user agreement from a Twitter licensor from May 2016 includes the restriction, but applies it only to law enforcement.

Moreover, it appears that concerns over profiling users' "political affiliation or beliefs" are part of a larger set of what Twitter has called "sensitive categories" since at least 2014, including "health, negative financial status or condition, political affiliation or beliefs, racial or ethnic origin, religious or philosophical affiliation or beliefs, sex life and trade union membership." The company prohibits these categories from being used for keyword-targeting advertisements, deeming such profiling "inappropriate or offensive" and that such use could potentially "compromis[e] users' trust."

A Twitter spokesperson subsequently confirmed that its policy did explicitly prohibit the use of Twitter data to identify or estimate the political affiliation or beliefs of any Twitter user as outlined in the paper and that the authors did not have an exemption to the policy for their study and pointed me to Twitter's research policy, which states "Please note that any use of Twitter data,

including for research purposes, remains subject to all parts of the Developer Policy and Agreement." The company, however, did not respond when asked whether it takes any action against violations of its policies, including educating researchers about their obligations under these agreements, leaving open the question of what purpose such policies have if companies do not actively enforce them with real consequences for violations by academic researchers (in contrast to commercial violations, which are dealt with strictly).

Given the submission dates of the conference workshop the study was presented at, it is certainly possible that the authors had completed their study prior to Twitter prohibiting such research in its main Developer Agreement and a Twitter spokesperson further clarified that it would depend on whether the authors actually accessed and analyzed the Twitter data themselves or whether they made use of derivative data or third-party analyses that had been previously performed by others, given that the paper does not clarify how the authors accessed the Twitter follower data or who performed what analysis. Yet, at the very least, given that Twitter had deemed political affiliation research to be a sensitive category for several years and that if the research was performed in the past 12 months a basic web search would have turned up growing prohibitions on such research, the authors would have known that such research was something Twitter was increasingly concerned with.

After two days of no response from the paper authors I followed up with them again to see if they could at the very least share whatever replication data they had available at the moment. I also asked if they could send me a copy of their final university IRB approval that outlined how they and their university IRB considered the question of researching a topic that Twitter considered extremely sensitive and had been increasingly prohibiting research of on ethical grounds.

Nearly a week and a half later, I have to date never received any response from the authors regarding any of my queries. Not one of the questions I asked about their methodology or algorithms have been answered and the authors have not provided any data that would enable me to conduct even a cursory spot check of their data and methods.

In short, I simply cannot replicate or validate any aspect of the study, short of assembling my own list of journalists publishing at each of the 25 outlets they identify, downloading the half million articles for those authors listed on MuckRack (which may itself be impermissible under that site's terms of use), selecting algorithms I believe would work in each case and crossing my fingers that with enough trial and error I can get some kind of workable result. Yet, without the ideological terms list which the authors manually adjusted using domain expertise or insight into which extraction and preparatory algorithms they used, it is impossible for me to replicate the authors' specific analysis.

Most importantly, it means I cannot perform the kind of cursory validity check that is the most common form of basic validation – checking, for example, to see if their list of ideologically weighted terms matches those compiled by other projects or seeing whether the list of 500,000 articles they downloaded are

relatively comprehensive of the coverage published by each author. Instead, any attempt at replication would have to repeat the entire massive analysis blindly and from scratch, rather than spot checking specific pieces of it - setting a bar far in excess of what most researchers could do.

Intriguingly, one of the paper's authors was in fact the lead author on a Policy Forum piece in Science just three years ago criticizing some of the public big data research coming out of the commercial world for not providing these exact replication datasets and algorithmic insight to allow others to replicate and validate their work. The authors of that piece noted "replication is a growing concern across the academy" and they single out one study as "not meet[ing] emerging community standards" because "neither were core search terms identified nor [the] larger search corpus provided" and that at the very least the study could have published aggregate data because there are "no such [ethical, legal or privacy] constraint[s] regarding the derivative, aggregated data" for that particular study.

In short, the authors of that 2014 Science piece criticize an industry paper as "not meet[ing] emerging community standards" for failing to publish the list of search terms it relied upon, yet three years later its lead author has published a paper that similarly does not include the list of search terms and does not respond to multiple requests to provide them. Indeed, the 2014 article notes that the example it focuses on offers "a case study where we can learn critical lessons as we move forward in the age of big data analysis," yet those lessons appear not to have gained much traction.

After failing to receive a response from the authors in my quest to obtain replication data, I reached out the organizers of the conference workshop the paper was presented at, hoping that perhaps the organizers could help facilitate access. One of the organizers responded on behalf of the workshop noting that neither the host conference nor the overarching professional society ACM require replication datasets be made available and that thus none were available for their workshop.

She also clarified that the workshop did not require authors to provide any evidence that their research underwent ethical review. When asked why this particular workshop did not request IRB approval of submissions, the organizers declined to respond further beyond noting that to their knowledge only one computer science conference (also under ACM) requires such evidence. ACM also did not respond to requests for comment on whether this statement was correct and if so, why it does not request any evidence of IRB approval for submissions to its conferences. Thus, the conference workshop itself was unable to assist in locating replication datasets or providing more information on how the authors addressed Twitter's inclusion of the research topic in its sensitive categories list.

Of course, as I've noted before, computer science is one of the disciplines that does not have a longstanding tradition of formal ethical prereview of research. Thus, while there are a variety of reasons that computer science outlets have become one of the dominant venues for "big data" research, ranging from their

rapid publication times to their familiarity with the complex techniques of data mining, they are also uniquely welcoming to big data research in that many do not impose the extensive ethical prereview processes that journals in other disciplines require and thus sidestep questions about terms of use agreements or questions about what kinds of research should be conducted.

A review of Northeastern's IRB website indicated that the university required IRB approval even when working with publicly available data, stating "All research activities must be reviewed by the Office of Human Subject Research Protection even when categorized as 'exempt' status" and repeating two sentences later in all capital letters "FINAL DETERMINATION OF EXEMPT, EXPEDITED AND FULL COMMITTEE STATUS IS MADE BY THE INSTITUTIONAL REVIEW BOARD."

This meant that even if the study had been considered under university IRB policy to be "exempt" (meaning it was not subject to a more in-depth IRB review), the IRB would still have had to review it and consider any ethical questions. The university did not respond to a request for comment on whether the study in question did in fact undergo IRB approval, but since university policy explicitly requires it without exception, it can be reasonably assumed that the university's IRB office did review and approve the research before it commenced.

If the research was conducted after Twitter had added "political affiliation or beliefs" to the list of prohibited categories of profiling, the IRB would have had to weigh the benefits of adhering to those requirements against the detriment of blocking research it thought would have value to society. Alternatively, if the research was conducted prior to it being added as an explicit prohibition for all users, the IRB would have had to weigh both its inclusion in Twitter's "sensitive categories" list and the company's rapidly growing list of user classes banned from using it, as evidence of the company's concern about the ethical considerations of profiling users by those characteristics and its belief that such profiling was ethically fraught.

Either scenario offers a fascinating glimpse into the complex ethical decisions confronting research ethicists in the big data era and a unique inversion of the typical storyline when it comes to big data research ethics. Most often it is companies performing research that the academic community later condemns as a violation of accepted ethical norms. Yet, in this case it is a commercial entity that has laid out a set of ethical guidelines for the use of its data where it has highlighted topics it believes are especially sensitive and ultimately explicitly prohibited research that profiles users along those dimensions as being outside what the company considers to be acceptable ethical standards. In contrast, a university IRB has, at the very least, determined that despite the company's designation of these dimensions as "sensitive categories," that the benefits of profiling users along those dimensions outweighs other concerns.

However, neither the researchers nor the university responded to repeated requests for more information on how this determination was reached or any details about the policies, evidence and normative standards used.

The university also did not respond to a request for comment on whether it has any institutional policies that require its researchers to make replication datasets available within the bounds of legal or ethical restrictions or whether it requires them to answer basic questions from other scholars attempting to replicate their work.

Given that one of the authors is also on the Board of Reviewing Editors (BoRE) of Science, I asked AAAS if it requires its BoRE members to make replication datasets available for their work or at the very least respond to reasonable questions regarding replication of their work.

A spokesperson pointed to the journals' publication policies, which require that for all materials published in Science, "after publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data or materials must be fulfilled. Any restrictions on the availability of data, codes, or materials, including fees and restrictions on original data obtained from other sources must be disclosed to the editors as must any Material Transfer Agreements (MTAs) pertaining to data or materials used or produced in this research, that place constraints on providing these data or materials. ... Unreasonable restrictions on data or material availability may preclude publication."

When asked if these policies applied to BoRE members publishing in journals other than Science, as a condition of their membership in BoRE and their stature as prominent emissaries of the profession, the spokesperson clarified that in such cases "they are subject to the policies of that journal although they are, of course, familiar with the spirit of our policies." The journal declined to comment, however, on any actions it might take to encourage or compel its BoRE members to cooperate with replication requests in those cases.

Thus, at the end of more than a week and a half of extensive research and correspondence with the authors, their university, the conference workshop organizers, the conference's sponsoring society, Twitter and AAAS, I have neither replication data nor answers to any of my questions in hand, nor any remaining avenues to pursue. I have no way to test whether inclusion/exclusion of quoted phrases or the selection of document extraction algorithm would impact their results nor can I evaluate how comprehensive or representative the list of reporters or coverage for each of those reporters was or whether their domain expert adjustments may have biased their findings. Short of coming up with my own study from scratch and investing the immense resources in performing my own complete study, I have no ability to confirm or refute any of their findings.

Yet, even if the authors did eventually someday provide replication data, it is unclear whether such replication could even proceed in light of Twitter's outright ban on the specific topic at the heart of the study.

If other researchers in the future attempt to research similar questions or are able to locate replication data for the study, their IRB will be confronted with

the ever more common question of whether academic researchers should be allowed to ignore the terms of use and other legal and licensing agreements that govern their access to the data that is the lifeblood of their research. As I've found over the last two years, such violations are increasingly commonplace across today's academic environs, with many IRBs appearing to accept violations of such agreements as allowable. In the case of researchers of the future attempting to replicate or expand upon this study, that future IRB will have to weigh its own scholars explicitly and knowingly violating Twitter's Developer Agreement with the benefits of being able to replicate or extend existing literature. However, this sets a course down an ethically and legally slippery slope: if it is ok to violate Twitter's binding legal agreement in the interests in science, would it similarly be ok for researchers to violate the university's agreement with Proquest and bulk download hundreds of millions of articles from the university's subscription and redistribute them to collaborators at other campuses without subscriptions, if this would enable an important project that could not otherwise be completed? Where is the line drawn?

Given that in this case Twitter's new prohibition on this research relates specifically to the company's determination that profiling along these dimensions would be a violation of acceptable ethical standards, that future IRB would be weighing not only a violation of a terms of service agreement, but specifically a violation of a prohibition put into place to protect human subjects from undue harm. If an IRB believes it is acceptable to knowingly violate such prohibitions in the interest of science, this raises the question of where it would draw the line when it comes to other ethical terms of use prohibitions. For example, take a deidentified medical dataset that explicitly prohibits reidentification of users – could researchers reidentify those users and combine their records with other datasets in explicit and knowing violation of those terms of use, if they believed this would help them answer questions they felt were important, but which the dataset owner believed were ethically unsound?

In short, as companies like Twitter place ethical use restrictions on their data which prohibit certain kinds of research on the basis of that research being determined by the company's ethicists to be unsound and a violation of accepted human subject protections, university IRBs will increasingly find themselves having to, quite literally, weigh conflicting standards of just how much protection human subjects should receive in a big data world.

What happens when those studies of the future are submitted to a journal where the author's university IRB has approved the research as ethical, but the dataset owner files a formal protest that its own IRB has ruled the work unethical and a direct violation of the human subjects protections section of the legal agreement governing the use of its data? When I posed this question to Science, it said in such a case it may seek additional outside ethical guidance, but when asked if it outright banned such papers or if it would automatically retract without question a paper that was approved by a university IRB but where the dataset owner in question requested retraction on the grounds that its own IRB had deemed the work a violation of ethical standards and a violation of the author's legal agreement to use the data, Science declined to say

it would blanket ban or retract such papers as a matter of policy and that it would instead depend on the specifics of the situation. This itself is a fascinating statement that one of the scientific community's flagship journals is unwilling to place a blanket ban on publications that knowingly and explicitly violate the legal and ethical agreements governing large datasets, especially highly sensitive topics like the reidentification of deidentified medical data, as long as the researchers' home institution IRB approved the work and even if other IRBs disagree and that it would instead depending on the specific circumstances.

Putting this all together, it is clear that the "replication crisis" doesn't simply refer to the inability of one set of scholars to reach the same conclusion as another set of scholars given the same dataset, but rather the inability to even acquire the datasets or basic answers needed to even attempt replication. Even for papers published by prominent faculty, it can be impossible to acquire replication data or get basic questions answered about the work to enable even cursory replication. In this case, after a week and a half of exhaustively attempting to track down replication data for the paper in question and contacting 19 people at 7 different institutions, I was forced to finally give up and accept that there is no way to replicate this study short of launching my own study from scratch. How many other scholars run into this issue every day when attempting to replicate or build upon the work of others in their field? In a world where even basic facts are now matters of dispute, how can we, as big data researchers, defend the output of our discipline if we do not, as one of the author's coauthored 2014 pieces put it, "learn critical lessons as we move forward in the age of big data analysis" regarding making replication more accessible?

Perhaps if there was more incentive to encourage the big data community to assist in replication as much as possible, whether by virtue of more conferences and journals enforcing the kind of stringent replication standards of Science, more universities placing mandatory data deposit requirements on their researchers (where permitted by legal and ethical standards), more professional societies and major boards like BoRE placing requirements on their members to reasonably assist in replication requests or leading scholars setting examples by providing replication data for all of their own studies, perhaps we might see greater interest in the topic, rather than it all-too-often being viewed as merely a bureaucratic burden. Similarly, as the diversity and sensitivity of available datasets expands and as dataset owners increasingly set greater restrictions on how their data can be used, especially ethical constraints prohibiting research they believe would cause unacceptable harm to human subjects, university IRBs will find themselves navigating increasingly perilous ethical and legal landscapes as they try to facilitate the infinite creativity of big data researchers while balancing those ideas against the rights of those being studied.

Only time will tell how this evolving landscape turns out, but in a world where our every action, from walking down a public street to our most intimate moments in our homes or doctor's offices are increasingly captured in datasets ready to be mined and where the results can influence public policy or even law, the stakes have never been higher.

Based in Washington, DC, I founded my first internet startup the year after the Mosaic web browser debuted, while still in eighth grade, and have spent the last 20 years working to reimagine how we use data to understand the world around us at scales and in ways never before... **MORE**