Democracy's Data Infrastructure: The Entanglement of Politics and Science

by Dan Bouk and danah boyd
4S Talk / August 20, 2020

Every decade since 1790, the United States has conducted a census. This laborious process produces the official count of the population residing in the country. This data is then used to apportion the House of Representatives, allocate federal funding, and serve as the numerical basis for the redistricting of federal, state, and local electoral districts. As many of our peers in the 4S community have consistently shown, this purportedly neutral act of counting heads is political all the way down. From the Constitutional logics of counting enslaved people as 3/5 of a person to the contemporary debates over citizenry, racism, eugenics, and nativism have been entangled in the count's procedures just as they've been enmeshed within the logics of the United States as a whole. Yet, time and time again, scientists have proposed new technical solutions to eliminate politics from the decennial count in ways that have consistently backfired. Our talk today is the story of two such attempts, separated by a century of time but sharing so many similarities as to be an eerie warning of what is about to unfold.

The Past

Our story begins in December 1920 when Joseph Hill, the Chief Statistician for the US Census Bureau provided the US House of Representatives with statistical tables printed for possible apportionments of Houses ranging in size from 435 to 483, based on "preliminary population figures, subject to revision. Hill chose to provide these data using an apportionment algorithm known as "major fractions." This was not an arbitrary decision.

After every census since 1790, the census figures were delivered to Congress who would then decide how to apportion the number of seats in the House to ensure that each state was fairly represented. Inevitably, Congress would increase the size of the House of Representatives such that no state would lose representatives. The method used to calculate the apportionment and divvy up the representatives changed over time.

After the 1910 census, Congress had decided that the House of Representatives should not keep growing in size but be constrained to 435 seats. Such declarations had happened before, but Congress typically ignored them and increased the size anyhow. It was easier to deal with conflicts through adding seats than negotiating out the zero-sum calculation that such a steady state would require. This is why Hill prepared calculations with different House sizes in mind.

Hill's decision to consistently use the "major fractions" algorithm in his recommendations was not based on his own preference. In fact, he had much preferred a different approach to allocating the mathematical remainders involved in dividing up House seats. But throughout the 1910s, a statistician named Walter Willcox had advocated for the "major fractions" approach, giving public lectures, advocating amongst scientists, and lobbying Congress and the President.

Hill was a bureaucrat; he was willing to accept and execute Congress' recommendations. And it appeared to him that Congress had settled on "major fractions" as the algorithm they intended to use.

Then, on the eve of Congress' debate in January 2021, a Harvard engineering professor named Edward V. Huntington called out the "injustice" caused by Willcox's method in a letter to the editors of the *New York Times*. Huntington championed a new method (which turned out to be Hill's method very slightly improved). While Huntington's new calculations only caused a change to the apportionment of two states in the bill Congress was considering, Huntington pressed a more fundamental issue: "it should be remembered," he wrote, "that what is really involved is a mathematical principle which admits of no gradation between truth and falsity." Like Willcox, Huntington asserted the primacy of theoretical and technical correctness.

While we can discuss the technical differences between these two approaches in the Q&A, Congress wasn't particularly interested in the scientific debate. Instead, they saw a hook. An excuse. A way to justify a decision that had long been in the making. Y'see, not everyone in the House of Representatives wanted to re-apportion Congress, whether by increasing the size of the House or not. A lot had happened demographically since 1910. Millions of immigrants – mostly from Catholic countries – had immigrated to the United States and were residing in a small number of quickly expanding cities. Plenty of Americans had also left rural parts of the country and moved to more rural settings. Southern Democrats, in particular, saw the writing on the wall; they did not want to lose their political power. While racism and nativism ungirded their goals, it was much easier to garner allies by arguing that the technical debates must be settled first. Conservative members now seized on Huntington's language of incorrectness. They pointed at the way Hill's earlier numbers had changed (as he switched from using preliminary figures to the actual counts). The practices of the scientists, and their fight to prove their own method to be *the* right one, created space for conservative officials to impugn the validity of the entire census and assert the impossibility of any fair apportionment. Eventually, the House passed a bill, using major fractions to apportion a House of 435 members, and sent it to the Senate. But the damage—to the legitimacy of the census—had been done.

The controversy over method seems to have provided a convenient pretext in the Senate, which by design accorded greater power to less populous states. The Senate killed the apportionment bill—quietly. The debates would continue throughout the 1920s. The United States never did re-apportion Congress until after the 1930 census. And only then after a Senator from Michigan settled the methodological controversy with politics, not science.

The Present

Fast forward to the present census, currently underway. Today's census is rife with challenges. While over 62% of all households have self-responded, there are now hundreds of thousands of enumerators going door-to-door to get information from people who have not. There's a lot that gets in the way of this being a seamless process. First, there's a pandemic. Who wants strangers knocking on their door? Second, there's fear. We have an Administration who is hellbent on ensuring that only *some* people are counted in the census. The racism and nativism of today have echoes in history, mirroring the debates surrounding the 1920 census. In 2018, the President tried

to shore up fear by demanding that the Census Bureau collect data on citizenship; this was shot down by the Supreme Court last year. A month ago, the President produced a memorandum that demanded that the Census Bureau provide a count of undocumented individuals living in the country that he could use for reapportionment. Most census experts believe he will use this data to alter the apportionment numbers. While legal challenges are inevitable, the fear is pervasive.

Amidst this highly partisan spectacle, there is also a technocratic debate underway. Long before this Administration came to power, the Census Bureau's technical experts had begun to worry about their ability to promise confidentiality to data subjects, a promise that had both legal and methodological implications. Title 13 of the US Code requires that Census Bureau employees ensure that no data collected for the census can be reidentified; they may only publish statistics. Moreover, ongoing research has consistently shown that the most vulnerable people in the country won't participate if the data is not kept confidential. But advances in mathematical techniques and the widespread availability of commercial data has made it possible to undo the confidentiality protections that the Census Bureau has relied on in the past. In short, it is now possible to take statistical tabulations and reconstruct individual records from those data. This can then be matched against other datasets to undo the confidentiality of census data. The Census Bureau felt as though they had to act. Starting in 2008, they began exploring a disclosure avoidance technique known as differential privacy in order to protect the confidentiality of the data. By 2016, they were moving forward with the goal of implementing this at scale for the 2020 census.

Today's Chief Scientist of the Census Bureau is John Abowd; he is the bureaucratic equivalent of Joseph Hill. Like his predecessor, he has seen his technical interventions as tools to de-politicize and de-bias the census. Yet, also like his predecessor, he underestimated how spectacle might undo both his technical and bureaucratic efforts. Abowd is an economist surrounded by computer scientists. He sees transparency as central to modernizing the census. He took it upon himself to help people understand all the ways in which data had always been made, not found. He wanted data users to understand the limits to census data, the way data had been altered in order to ensure confidentiality, and how data users' obsession with small-area geography had resulted in data that were statistically wobbly. No one wanted to hear this. In his mind, modernizing the disclosure avoidance system would ensure that that data could be protected and that statistical work would be valid, provided that data users used the data appropriately. He saw transparency as fundamentally beneficial because it would allow data users to understand the limits of the data they were using. That has not been how this change has been perceived.

Making disclosure avoidance more visible made it more open to attack. Data users, it turned out, weren't interested in technical models; they were reasonably interested solely in what the system did to the data they cared deeply about. And what they saw in demonstration products outraged them. The data appeared wrong. Moreover, they felt excluded from decision-making. The communications efforts, when they did take place, resulted in more animosity than bridge-building. Some data users accused Abowd of destroying their data for his "science project."

The Census Bureau is now stuck between a proverbial rock and a hard place. Data users are demanding that the data be published without the newfangled disclosure avoidance systems in place. To do so, the Census Bureau must decide between radically limiting what data they make

available or publishing the data in ways that can be easily reidentified. Data users want them to just use the previous methods, arguing those were good enough. Yet, internal assessments suggest that scaling up the previous methods would result in deeply biased and unusable data. Many data users do not believe that the threat to confidentiality is real and reject the options given to them as false choices.

As the Census Bureau chugs along with its effort to count everyone, the stage is being set for the census to come undone. The data's legitimacy are being questioned from all sides. Conservative politicians are working hard to prevent a complete count and ensure that editing techniques that are used to improve the data are considered taboo. The Administration is waging war against the Constitution in their effort to re-interpret what it means to count everyone. And scientists are arguing amongst themselves about whether they could even trust data that had been altered to ensure confidentiality. In some way, the writing is on the wall; the real question is: what will be used to undo the legitimacy of this census? And what will be the ramifications of this?

STS in Situ

Among this community, we all know that data are made. Like the processes that make them, the data are socially constructed. We know that this is political work, even if it's not always partisan work. Yet, we also know that apolitical fantasies have power. They allow us to imagine that we can have sensible data infrastructure, that we can make meaningful data-driven decisions. The undoing of these fantasies doesn't necessarily result in a more informed public – or a more informed expert community. The undoing of apolitical data fantasies is itself political work, the political work that is at the root of what Robert Proctor, Ian Boal, and Londa Schiebinger all call agnotology, or the study of ignorance. While seeding doubt is a critical component of doing science, it can also be weaponized to undermine the legitimacy of institutions, the validity of data. While scientific debate may enhance knowledge, it can also be a political tool.

As we sit here, in the midst of a political spectacle, eyeing a Constitutional crisis on the horizon with our STS glasses on, it's hard not to cringe. While the partisan politics and resultant outcome may not be clear, what is clear is that the very tools of knowledge that we rely on to understand the world are being politicized in front of our very eyes.