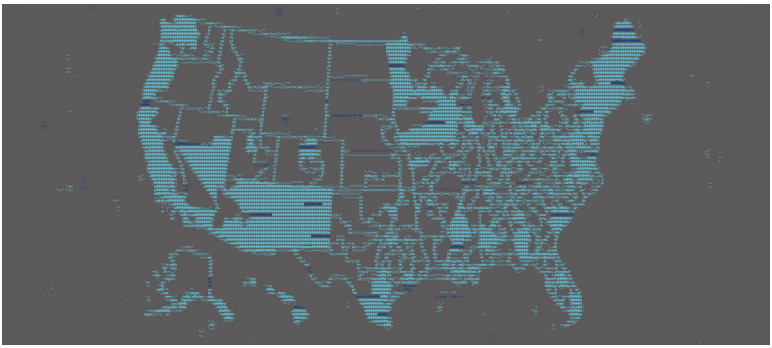
DATA AND DEMOCRACY



Democracy's Data Infrastructure

By Dan Bouk and danah boyd

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

In October 2020, the Knight First Amendment Institute at Columbia University convened a virtual symposium, titled "Data and Democracy," to investigate how technological advances relating to the collection, analysis, and manipulation of data are affecting democratic processes, and how the law must adapt to ensure the conditions for self-government. This symposium was organized by the Institute's 2019-2020 Senior Visiting Research Scholar, Yale Law Professor Amy Kapczynski, and co-sponsored by the Law and Political Economy Project at Yale Law School.

The essays in this series were originally presented and discussed at this two-day event. Written by scholars and experts in law, computer science, information studies, political science, and other disciplines, the essays focus on three areas that are both central to democratic governance and directly affected by advancing technologies and ever-increasing data collection: 1) public opinion formation and access to information; 2) the formation and exercise of public power; and 3) the political economy of data.

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, Executive Director; Katy Glenn Bass, Research Director; Amy Kapczynski, Senior Visiting Research Scholar; Alex Abdo, Litigation Director; and Larry Siems, Chief of Staff. The essay series was edited by Glenn Bass with additional support from Lorraine Kenny, Communications Director; A. Adam Glenn, Writer/Editor; and Madeline Wood, Communications and Research Coordinator.

The full series is available at knightcolumbia.org/research/

INTRODUCTION

In July 2018, the U.S. Census Bureau issued a Federal Register notice asking census data users to provide information about what data they used, at what level of geography, and how these use cases might affect different populations. This request for feedback did not announce to those users—local governments, social scientists, and the public at large—that its purpose was to facilitate a massive change in how data about the nation would be produced. A month later, the bureau's chief scientist, John Abowd, announced on a government blog the need to "modernize" the system used for disclosure avoidance by turning to "formal privacy" methods. Such an approach would allow the government to renew its guarantee of the confidentiality of data. The description of this innovation did not convey the trade-off that would be discussed later, a trade-off that underpinned the Federal Register notice, a trade-off that pitted accuracy against confidentiality, and then pitted both against the desire for conventional statistical tables with numbers that appeared to resemble counts. Unaware of the broader context, the solicitation for feedback confounded data users and community advocates. Which data mattered? All of it, they answered. They needed *all* of the data, and it had to be accurate. As the reason for this

question became clearer, data users grew upset, sides formed, coalitions coalesced, a controversy bloomed.

Knowing what we know about the challenges the 2020 census has faced, the controversy stirred by the Federal Register notice and Abowd's announcement of the new "disclosure avoidance" system may seem like a tiny technocratic matter. Even in 2018, another announcement loomed much larger. In March of that year, the Secretary of Commerce Wilbur Ross had announced that the decennial census would add a question about the citizenship status of every respondent. Civil rights advocates, state officials, and professional statisticians all protested. They argued that the question would likely cause significant undercounts of Latinx individuals, as well as immigrants and their relatives. A surprise Supreme Court decision in June 2019 disallowed the question, but a little over a year later the Trump administration came back with new plans to discourage participation in immigrant communities or even remove some immigrants from the count (despite a constitutional requirement that all residents be counted). These plans added confusion and uncertainty to a situation already mired in both. Amidst these political controversies, the Census Bureau encountered a new adversary that undermined a decade of planning and precise scheduling: COVID-19. As the Census Bureau struggled to complete the count amid a pandemic, a sea of bureaucratic confusion whipped up by the Trump administration destabilized its schedule. Temporal buffers were squeezed and squashed, statutory deadlines were missed, and court battles are ongoing. The legitimacy of the 2020 census, now the responsibility of the incoming Biden team, remains in peril. In the context of such extraordinary drama, one could be forgiven for thinking that a technical debate over confidentiality and accuracy might fade into the background.

And, indeed, it still might.

Or it could be that disclosure avoidance will become a very big deal.

We set out—an ethnographer and a historian—to think about the possible significance of the debate over disclosure avoidance by examining deeply the history of similar technical disputes that had troubled the census. We hoped our investigation might help us better understand what was at stake in the debates we were witnessing. More than that, we hoped that our investigation would teach us something about the role of openness in the making of open data.

The Census Bureau has long been one of the most consistent champions of the sorts of practices now marshalled behind the "open data" banner.[1] And, in theory, the new disclosure avoidance system sought to maintain that openness by allowing fine-grained data to be published, while still protecting the privacy of individuals. The technique employed for avoiding disclosures, called differential privacy, had been and still is championed as a tool for opening previously locked data sets—data held by private corporations or other bureaucracies that couldn't otherwise be revealed. Yet, in the case of the census, where a greater number of statistics had been published each subsequent decade, deploying differential privacy would appear to be a constraint. To data users, differential privacy came to signal a reversal of openness, both in terms of process and the data itself. Moreover, the Census Bureau's efforts at transparency were viewed as the opposite by those who could not make sense of either the code or the stilted technical communications. Transparency, it would seem, did not always strengthen trust.[2]

Like all infrastructures, the U.S. decennial census typically lives in the obscurity afforded by technical complexity. It goes unnoticed outside of the small group of people who take pride in being called "census nerds." It rumbles on, essentially invisible even to those who are counted. (Every 10 years, scores of people who answered the census forget they have done so and then insist that the count must have been plagued by errors since it had missed them, even though it had not.) Almost no one notices the processes that produce census data—unless something goes terribly wrong. Susan Leigh Star and Karen Ruhleder argue that this is a defining aspect of infrastructure: it "becomes visible upon breakdown."[3] In this paper, we unspool the stories of some technical disputes that have from time to time made visible the guts of the census infrastructure and consider some techniques that have been employed to maintain the illusion of a simple, certain count.

In ordinary times, the technocrats who run the census, the specialized staff that oversees the data, the army of enumerators who gather data, and the varied group of politicians, jurists, and scholars who use census data all do the crucial, seldom-acknowledged labor of hiding the seams in the census infrastructure.[4] This is the flipside of Star and Ruhleder's observation that breakdown reveals infrastructure—during ordinary times of operation, the community of practice that relies on an infrastructure maintains a regime

of infrastructural invisibility. All of their practices say, in effect: *nothing to see here*. Infrastructures, then, depend on the intentional production of ignorance. As scholars invested in agnotology, or the study of ignorance, we must then untangle the social, political, and technical mechanisms that shroud democracy's data infrastructure in opaque, and productive, knots.[5]

Technical disputes temporarily pierce those shadows. In the midst of such disputes, the community of practice speaks aloud assumptions that had previously gone unvoiced, and calls into question ideas typically taken for granted. For those same reasons, controversies have long been the primary objects studied by Science and Technology Studies (STS) scholars.[6] Controversies open the infrastructure to scrutiny, but not just to scholarly examination. So long as no one else is looking, technical disputes flare and then burn themselves out, resolving this way or that. But when the controversy breaks at a moment when others *are* looking—when, for instance, a global pandemic and political firestorm have drawn attention to the census—then the revelations made possible by a technical dispute can have unexpected, and powerful, consequences.

Whether anyone notices it or not, the decennial census matters a great deal. It is a massive mobilization of people and resources, like little else that happens in peacetime. Its goal is to count each person so that each will be represented in government and be supported by government programs. It generates facts about the nation that are widely accepted and the basis for legislation, planning, and debate. That is why we refer to it as the data infrastructure of American democracy.

Opportunities for conflict over the census and its operations abound. That's because there's nothing simple about the project of counting every single American. Since it was first conducted in 1790, the very act of operationalizing a count of all people has opened a Pandora's box full of questions, contradictions, competing values, political interests, and statistical uncertainties.[7] Who should be counted? Who can be trusted to do the counting? How do you count people who don't want to be counted? Where should someone be counted? When should the count happen? How do you know you're done counting? These are but a few of the questions that must be answered by a census operation. The more one sits with them, the easier it is to see the census as a large sociotechnical system shaped by its peculiar

history.[8] Moments of technical controversy produce threads to tug at to see the larger weave, just as they create opportunities for those working the system to their own ends.

A FIGHT THAT SHOULDN'T HAVE MATTERED

ONE OF THE MOST FAMOUS technical controversies in census history set two scientists against one another a century ago in a public fight to define the best way for Congress to use the data that the 1920 census produced. The immediate subject of their quarrel was the method of "apportionment"—also called "reapportionment"—the procedure that followed every census, where each state in the union was allotted a certain number of members of the House of Representatives in proportion to that state's population. One of the quarreling scientists, the economist Walter Willcox, dreamed of a system that would get Congress out of the picture. He imagined a mechanized procedure that would translate census figures into a proper apportionment of seats using a logic and method that resembled that which Congress had employed in prior decades. Willcox's automatic system would emulate Congress's action, but without actually involving Congress. His rival, the engineer Edward V. Huntington, believed a different method—one never before used by Congress—was mathematically and scientifically superior and so it was unfair and improper for Congress to employ any of the methods that it had used before.

The opinions of two scientists, even two well-regarded scientists clothed in the prestige of Cornell and Harvard Universities, respectively, should not have mattered. What really mattered was what Congress could agree to. According to the U.S. Constitution, Congress was to apportion House seats subject to these constraints: that representation should be proportional to state population as determined by a census, that the average ratio of constituents to representatives should never be less than 30,000 to one, and that each state was guaranteed at least one representative. Beyond those criteria, Congress possessed (and possesses) full authority to decide the best way to apportion. Congress decided how many seats there should be in the House and then how many seats each state was due, a question that was made

complicated by the fact that the division never worked out perfectly. Some states were always due a fraction of a representative, however the calculation was done, so it fell to Congress to decide how to deal with those fractions. Apportionment bedeviled Congress—indeed, George Washington (and the nation's) first presidential veto had been of an apportionment bill. But Congress had always managed some kind of apportionment, usually by making the House of Representatives big enough so that few or no states would be forced to lose a seat that they already possessed. Congress had authority over apportionment and a long history of using it.

A fight between two scientists should not have mattered. But it did. In 1921, their fight coincided with, and became enmeshed in, a struggle between those intent on freezing the House at 435 seats and those intent on its continued expansion. At first, the scientists' quarrel provided fodder for the expansionists. In the process, it helped derail the entire apportionment of the House for the rest of the decade. As a result, rural, native-born whites in the Mississippi Valley held on to power in the House of Representatives that they would otherwise have lost had the apportionment taken place. When the scientists shut up, their silence made possible the installation of an automated apportionment system, one that effectively locked the size of the House in perpetuity (and so also limited the size of the Electoral College, further privileging less populated states in presidential elections). A fight between two scientists should not have mattered so much. But it did, and a closer look at the story of their fight can show us why.

Willcox revealed his dream of automating congressional apportionment in his December 1915 address as president of the American Economic Association.[9] For years, Willcox had carefully studied the entire history of congressional apportionment. He claimed to discover a series of criteria that Congress either explicitly or implicitly employed. Some, he said, were inferred from the Constitution, like the responsibility to ensure as "near as may be" that each person enjoyed the same share of representation as all others, or the importance of setting each state, whether its population was large or small, on an equal footing. Other criteria had developed over time in Congress, like its habit since 1840 of deciding that every time the calculations said a state deserved a half of a representative or more, the state would get that representative added to its total. (Otherwise, they would have to round

down, since the House did not allow a person to serve only as one-half of a member.) Today, we might look at Willcox's plan and say that he had developed an expert system. He had studied Congress's ways of judging every prior apportionment (this was its training data) and built its reasoning into a method that would achieve its goals automatically. He called it the "method of major fractions."

The first step of Willcox's plan had been achieved. In 1911, he convinced Congress to adopt his method for making its apportionment decisions for a House with 433 members (which would grow to 435 when Arizona and New Mexico became states). Now in 1915, he hoped to garner academic support affirming that "the method of major fractions is the correct and constitutional method of apportionment."[10] With his method firmly established, Willcox aimed to nudge Congress out of the picture. As he explained to his peers, "It is now possible for Congress to prescribe, in advance of an approaching census, how many members the House shall contain, to ask the Secretary of Commerce to prepare a table apportioning just that number in accordance with the method of major fractions, and to report the result to Congress or to announce it by executive proclamation."[11] This automated procedure would rationalize the process of apportionment, Willcox hoped, removing political horse-trading and negotiation from the process, while making it easier to prevent the House from continuing to grow. But Willcox did not specify how he would convince Congress to allow itself to be so bound and mechanized.

Willcox could not have imagined how it would come to pass that Congress would indeed take up his idea. First, the U.S. Census Bureau had to accomplish the always monumental task of counting the entire U.S. population, and, in 1920, it had to make that count shortly after a world war had ended, amidst a flu pandemic, under the shadow of riots that targeted the nation's Black residents, and in the context of widespread labor unrest rising to challenge powerful corporate monopolies. The difficulty of completing such a count is not, unfortunately, all that hard for us to imagine today.

The Census Bureau did manage to complete the count, but only after extended delays that meant it had to first submit preliminary population totals to Congress in December 1920. At that time, the chief statistician, Joseph Hill, turned over to Congress tables printed for possible apportionments, using Willcox's major fractions method for Houses ranging in size

from 435 to 483.[12] A month later, the Census Bureau revised those apportionment tables using final, official population statistics in time for the committee's formal report, issued on January 8, 1921, which supported an apportionment by major fractions of 483 members.[13]

On the eve of Congress's debate, Huntington called out the "injustice" caused by Willcox's method in a letter to the editors of The New York Times.[14] Huntington championed a new method, one he had unveiled before the American Mathematical Society a week earlier on December 28, 1920.[15] On closer examination, it turned out to be derived from a method devised a decade earlier by the Census Bureau's Hill, who subsequently became Huntington's ally. The new method was first called the "method of the geometric mean" and then "equal proportions," although it is often referred to today as "Huntington-Hill." It too dealt with the problem of assigning seats to states based on their populations and offered a way to deal with fractions. What the method of equal proportions promised was a solution such that if one divided the average congressional district's population of one state by any other, the resulting ratio would be the smallest it could possibly be. Huntington's method minimized the percentage difference in population size of average districts. From early on in the debate, everyone stipulated to the consequences of this approach: Huntington's method favored less populous states and Willcox's favored more populous states.

Huntington's calculations only caused a change to the apportionment of two states in the bill Congress was considering, but 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."[16] Huntington's public critique highlighted the question of method, a part of the machinery for apportionment that seldom endured much scrutiny, making it suddenly both visible and also apparently important. And that visibility came in the thick of tense political combat at a propitious moment.

Politicians in Congress used Huntington and his critiques for their own ends. The Republican conference controlled the House, but it split between those who favored keeping the House at only 435 members and those who wanted a much larger House of 483 members, at which size (not coincidentally) no states (and so no incumbents in Congress) would lose a seat. The

Census Committee chair had put forward a bill for 483 members, but the Republican leadership was succeeding in pushing through an amendment for a smaller House.[17] Huntington wrote another letter. This one, which was read aloud on the floor of the House, explained that for 435 seats his method produced different results for six states, not just for two. Seats granted by Willcox's calculation to New York, North Carolina, and Virginia should instead be awarded to New Mexico, Rhode Island, and Vermont, Huntington argued. Those pushing for a larger House could see they were losing and so they latched onto Huntington's objection as the pretext for a delay. That gambit failed at first—the House passed an apportionment for 435 members using major fractions.[18] But it worked in the long run. Huntington wrote another letter to the Senate Census Committee chairman begging to have his case heard.[19] The Senate—filled with members from states that were not happy about losing seats in a smaller House—obliged, withholding the bill from a vote under the pretext of needing to gather scientific opinions on the methods controversy.[20]

The methods controversy continued to serve as one of a handful of tools for preventing a 435-seat reapportionment throughout the 1920s. As a result, representative power remained in states that were more slowly growing, and which also happened to be home to more native-born whites than many other parts of the nation—this on the eve of the passage of a radical bill restricting immigration to the United States.

Yet for Willcox, Congress's failure to apportion—and the unconstitutional injustice entailed by that failure—presented an opportunity nonetheless. Beginning in 1926, the first of a series of bills passed through the census committee proposing a "ministerial" apportionment. The bills charged an official in the executive branch, the secretary of the commerce first and later the president, with the task of making an apportionment by some predetermined method for a predetermined size of the House (in this case, remaining at 435).[21] Willcox had been advocating this course of action for over a decade—automating Congress's power was an act of desperation for legislators from states deprived of their due representation (like California and Michigan), but it was what Willcox wanted all along.[22]

Huntington's method of equal proportions stood in the way. Huntington and his Census Bureau ally Hill rounded up statements of support from elite

statisticians and mathematicians, including the chairs of the math departments at MIT and Iowa, the dean of Michigan's business school, bank and life insurance executives, one former census director, and past presidents of the American Mathematical Society, American Economic Association, and American Statistical Association.[23] Huntington inundated legislators with letters and arguments, his style then and always being to demonstrate the soundness of his arguments through the volume of his writing. One promising bill began its life using Willcox's method of major fractions, but when it finally made it to the House on March 2, 1927, it had adopted Huntington and Hill's method of equal proportions.[24] The controversy over method helped sink that bill and those that followed.

Silencing the scientists proved a key strategy in finally passing an apportionment bill. The newly appointed senator from Michigan, a Grand Rapids newspaperman and progressive Republican named Arthur Vandenberg, needed to prevent the methodological controversy from providing fuel to his opponents—those who preferred to stymie yet another apportionment and so hold on to power. Vandenberg's state had been one of those denied due representation by the failure of the 1920 apportionment, and Vandenberg was determined to do whatever it took to get his state's four additional representatives (and Electoral College votes).[25] That meant removing an explicit reference preferring any particular apportionment method, instead just specifying that the last one used would be used again (giving the nod implicitly to "major fractions"). Finally, to get the bill passed, Vandenberg appended it to the 1930 census law, the law that enabled the coming census and with it the roughly 100,000 enumerating jobs that Vandenberg's colleagues in Congress were eager to distribute as patronage. To pass the bill granting access to that patronage, the Senate had to approve the automatic apportionment too.

Huntington continued to draw attention to the technical controversy, destabilizing the situation. As late as January 27, 1929, Huntington took to The New York Times to claim his mathematical principles were equivalent to the principle of apportionment itself. "In its anxiety to satisfy one provision of the Constitution, which requires a re-apportionment of the House of Representatives every ten years, Congress is in danger of overlooking another provision of the Constitution, which requires that the number of seats assigned

to each state shall be proportional to the population."[26] Vandenberg sought in a series of increasingly pleading letters to dissuade Huntington from such an equivalence, writing in one: "I am sure you will agree with me that the paramount need is to *get* re-apportionment—regardless of the method used. ... You can help incalculably in getting this enabling legislation. The greatest possible help would be your agreement with me upon this amended form which undertakes to leave the pending legislation entirely above and beyond any argument over methods."[27] Huntington could help "incalculably" by stopping his calculations, by saving his fight until after the ministerial apportionment bill had been passed. Vandenberg convinced Hill at the Census Bureau to help dissuade Huntington from keeping up the fight.[28]

Huntington appears to have quieted what Vandenberg called his "busy Harvard mimeograph" long enough for the ministerial legislation to pass.[29] But Huntington was stubborn that science had nothing to do with politics: "It is inconceivable to me that any proper legislation can be hampered in any way by such impersonal and colorless things as correct mathematical facts, and I decline to be classed as an obstructionist, either wittingly or unwittingly, of your legislative program," he told Vandenberg.[30] After which, he went right back to politicking for his method of equal proportions, working even more closely over the coming decade with the Census Bureau.

In the end, equal proportions did win, but not primarily because of the power of Huntington's arguments. The final bill that had passed in 1929 provided that at each apportionment, the Census Bureau would print the total state populations and the number of seats due each state, according to both the method of major fractions and Huntington and Hill's equal proportions. In 1930, major fractions controlled the apportionment, since it had been used last—but the method turned out not to matter, since both methods happened to agree at a 435-seat House. In 1940, the methods differed, but on only one pair of states. That was enough to doom "major fractions." The method of equal proportions took a seat away from Vandenberg's own, reliably Republican Michigan and handed it to the solidly Democratic Arkansas. Vandenberg fumed, but the bill passed, and Huntington and Hill's method became the default method anchoring a system of automatic apportionment that has now operated without interruption for nearly a century.[31]

A fight between two scientists should not matter all that much when

we're talking about the practice of Constitutional duties with century-old precedents. But the fight between Willcox and Huntington did matter. Their dispute shed light on a complicated problem that political actors could use to their advantage. It became a useful pretext for delaying early attempts to reapportion the Congress in 1921 at 435 seats and in later years threatened to prevent further reapportionments. As result, both scientists became unwitting accomplices to an unjust decade-long transfer of power away from more urban states that welcomed immigrants. Then, when all was said and done, after two decades of fighting, Willcox and Huntington each won something they wanted—Huntington won the use of his preferred method (which favored less populous states), and Willcox saw the fulfillment of his dream for a system that sidelined Congress and allowed both the House and the Electoral College to grow less democratic with every decade. Arguably, these victories were an even greater injustice than the earlier failure of apportionment.

THE MESS THE UNDERCOUNT OBSCURES

ANOTHER SET OF famous census controversies involved the so-called "undercount," a measure of the percentage of the population missed by the census. These late-20th-century controversies tore holes through a veil that had been intentionally woven to obscure perpetual problems with the census. To understand the effect of those controversies, we must first understand where the undercount came from and its role in assuring the invisibility of democracy's data infrastructure.

The census undercount was quantified in 1899 by the same Walter Willcox of the "apportionment" battle described above. He made his calculations in the context of a chorus of professional social scientists campaigning for the reforms that would eventually lead to a perpetual, independent census office. Willcox wanted to determine how accurate the current decennial census counts were, so he compared those counts to similar enumerations made by the individual states. Based on his findings, Willcox decided that "the last federal enumeration was probably within one per cent of the truth." He continued, "I believe that the faults of census legislation and administration

have impaired public confidence in the results here considered more than the facts warrant."[32] The problem was not the result of the census, but rather the optics of its operation.

Census operations were often messy—some people always got missed and some people who were counted claimed that they hadn't been. Census operations in the late-19th and early-20th centuries graduated not infrequently from messy to scandalous. Every 10 years, big cities charged that they had been undercounted. Cities also perpetrated systematic fraud. In the 1890 census, Minneapolis's enumerators engaged in rampant "padding," filling in tens of thousands of names for people they had not actually counted, and got caught by the Census Bureau.[33] These were the sorts of stories that impaired public confidence.

The interesting thing about Willcox's error calculation is that it was explicitly intended to play down padding scandals and controversies that led to recounts. This, as we'll see, was (and is) the norm. Calculations and investigations of error worked to quiet outrage and protect the infrastructure from greater or potentially damaging investigations.

On the eve of the 1920 count, a small corps of statistical workers within the Census Bureau attempted to recast the undercount as a calculation that could serve the cause of racial justice. The Census Bureau published a report in 1918 called "Negro Population 1790-1915." Its chief credited author was a white Chicago-educated Ph.D. (in political economy)[34] named John Cummings, who worked as a "special agent" to the bureau in the recently segregated federal government.[35] But most of the analysis was actually done by a "a corps of Negro clerks working under the efficient direction of three men of their own race, namely, Robert A. Pelham, Charles E. Hall, and William Jennifer."[36] Their work discovered an "undercount" in 1870 that, on a national level, missed 10 percent of Blacks compared to just over 2 percent of whites. They further argued from the "improbability" and "inconsistency" of various census-reported rates of growth, birth, and mortality that the 1890 census had missed a very large portion of the African American population.[37] Put another way, their scientific evidence challenged Willcox's purportedly precise assessment that that same census was accurate to within 1 percent.

The report singled out the 1870 and 1890 censuses for their inaccuracies in counting Black Americans. But the authors also posited a more general

and persistent racial undercount, one that would be obscured by a single undercount number. "It is not improbable that at other recent censuses the proportion of omissions has been higher, and the proportion of duplications lower in the enumeration of the Negroes than it has been in the enumeration of the whites; and that in general the margin of error has been greater in the case of Negroes," they wrote.[38]

In 1922, the Black mathematician and public intellectual Kelly Miller extended the undercount critique in light of the returns of the 1920 census. Miller cited the 1918 Census Bureau report as evidence of the possibility of a race-specific undercount and then proceeded to make a case that Blacks had been undercounted again in the 1920 census. The crux of Miller's argument was that the 1920 census showed that the rate at which the African American population was growing had been cut in half since the 1910 census. Miller blamed this apparent decline on the difficulties of counting many young Black people who had moved to cities or headed North during World War I.[39] He also faulted the bureau's data for birth rates (since most of the South still relied on sub-par birth registration systems) and so refused to believe the bureau's claim that African Americans had simply had many fewer children in the last decade.[40] "It is particularly unfortunate," he wrote, "that such loose and unscientific propaganda can be bolstered up by data from governmental documents which the uninquiring mind is disposed to accept with the authority of holy writ. ... The thought, and perhaps the conduct, of the nation may be misled on the basis of erroneous data, backed up by governmental authority."[41] Miller's explicit goal in his article was to attack the legitimacy of census data so often used in debates over "the Negro problem."

The bureau responded two months later with a rebuttal to Miller. Le Verne Beales, a white Census Bureau special agent, refused to take responsibility for the errors of 1870 and 1890, or to let them impugn the 1920 results: The new, permanent Census Bureau did better, he claimed.[42] Beales picked apart pieces of Miller's argument and claimed by his own analysis that the 1920 count of African Americans seemed reasonable. Moreover, he wrote that Miller had failed to account for the lower birth rates that might come with the migrations of Black people from the countryside and did not account for excess deaths from the influenza pandemic. The prior decade had been "abnormal," he asserted, as his colleagues in the bureau and members of

Congress had been saying since well before the 1920 count even began.[43] From his vantage point, that abnormality, and not Census Bureau incompetence, explained why the African American population seemed abnormally low. Beales finally returned to the idea of a single, small undercount. He found "no ground whatever for attacking the 1920 census as inaccurate beyond the small margin of error which is inherent in any great statistical undertaking."[44] This was the work of the undercount number, to admit some small amount of error while deflecting more radical critiques of the counting infrastructure.

Miller's was far from the most dangerous attack on the census infrastructure in the 1920s. Members of Congress frequently used the inaccuracy of that particular census as an excuse to justify denying an apportionment at 435 seats. Mississippi's John E. Rankin claimed in 1921 that he believed 10 percent of his state's population to have been missed.[45] Rankin occupied a seat on the House Committee on the Census and used that place for years to fend off all apportionment bills at 435 seats, grounding his resistance in the charge that the census was too inaccurate to justify taking away a seat from his state's delegation. Representative Henry Ellsworth Barbour of California called on the Census Bureau to repudiate Rankin and his doubts. Joseph Hill wrote a letter to Barbour, to be read aloud to the House, that said "we have no reason to believe that errors of this kind were any more frequent at the census of 1920 than at any other census, or were serious enough to vitiate in any appreciable degree the substantial accuracy of the results."[46]

The passage of the 1929 automatic apportionment legislation appears to have made it even more important that the census data be defended and defensible. In prior years, the final responsibility for judging the data and its fitness for apportionment had fallen to Congress. But the new automatic system appeared to make the Census Bureau the final arbiter. As the Census Bureau's director put it in a 1929 letter, "Of course you will recognize the necessity of making an enumeration of the population that cannot be successfully criticized. The operation of the new apportionment law will be based upon this census. The enumeration for the census of 1920 has been severely criticized."[47] That pressure to avoid criticism inspired new measures to quantify quality.

A statistical movement driven by probabilistic methods offered one path

forward. In the 1930s, a probabilistic revolution swept through the Federal Government, and the Census Bureau especially.[48] In 1937, the revolutionaries made their mark by turning what was supposed to just be a "check" of a larger self-registration system into *the answer* to the question, "How many Americans remain unemployed?" Statistical sampling allowed Census Bureau scientists to get their answer far quicker (and more reliably, they showed) than the more cumbersome process favored by President Roosevelt. They brought what they learned to the census in 1940, including for the first time a 5 percent sample in the decennial count.[49] One of the champions of sampling, Calvert Dedrick, also began thinking about how to more systematically check whether the census had successfully enumerated the entire population.[50] His plan would take another decade to come to fruition.

In 1950, the Census Bureau rolled out a "post-enumeration survey," or PES, that sent out highly trained enumerators to "recanvass" 3,500 randomly selected "small areas" and "reinterview" 22,000 randomly selected households. Post-enumeration meant that this well-trained band of social surveyors went out into the field after the initial complete count (the "enumeration") had been completed. They picked areas and households at random to count and interview them again. Then they compared the results to see how many errors the initial complete enumeration had made for those areas and households. That comparison and some fancy math allowed Census Bureau scientists to estimate the total extent of the undercount. The 1950 PES reported an underenumeration of at least 1.4 percent (with a standard error of 0.2 percent) for the entire U.S. population and 3.3 percent for the "nonwhite population."[51] Those were comforting numbers, not all that far away in total from Willcox's 1899 assertion.

The other path forward in assessing census accuracy took the name "demographic analysis." This was a method that grew out of the sort of analysis that Willcox, Cummings, Miller, and Beales had all performed. Demographic analysis begins with aggregate census data, grouped by age, race, geography, and other demographic characteristics. It uses data from administrative records to determine average rates for birth, death, and migration for each demographic group. The analyst then computes the expected population in a given year for each demographic group. Differences between the expected population and the actual count are presumably evidence of possible error.

In 1947, a researcher at the University of North Carolina named Daniel Price published evidence of an undercount from demographic analysis.[52] Price took advantage of a bureaucratic data set created by World War II. Six months after the 1940 Census began, the Selective Service Act required all men aged between 21 and 35 to register themselves. Those who did not faced a very steep penalty. Price compared the counts by state for this group of men and discovered a stunning failure to count African Americans. The net undercount of all Americans totaled about 3 percent of the population, which did not seem to worry Price overly much. But the net undercount of 15 percent of African Americans concerned him much more. At the end of his paper, Price revealed his own politics in a wry closing paragraph, featuring this line: "These results also make it seem that the advocates of 'white supremacy' in Mississippi might have been celebrating prematurely when the 1940 Census showed Mississippi for the first time had more whites than Negroes."[53] Black people, he informed his Mississippi readers, still constituted the majority in the state.

Price's paper is often invoked as the beginning of the undercount story, but it was not. It was not even, as Ken Prewitt has argued, the beginning of the racialization of the undercount.[54] A precisely quantified undercount had existed since at least 1899 and had been bound up with discussions of race for decades when Price published his study. Price's study also does not seem to have been particularly controversial. Rather it did what undercount quantifications usually did: It narrowed the problem of census accuracy to a single set of metrics and mostly served to support the claim that the decennial census counted very nearly everyone.

Measures of undercount constituted (and still constitute) an infrastructural adaptation meant to safeguard the legitimacy of the census. But in the late-20th century that adaptation itself became the locus of a controversy that extended over 30 years and multiple cases decided by the Supreme Court.

The Census Bureau's scientists wanted to get to the bottom of the undercount, to understand where it came from, so that they could design procedures to eliminate it. Census Bureau scientists, in an internal report in 1957, pointed out just how little the bureau knew about the true extent of over- and under-enumeration. They worried especially about "uneven distribution of under-count," and concluded that "major changes in procedure are

necessary to cover the reluctant and hard-to-find parts of the population."[55] After the 1960 census, researchers added to the post-enumeration survey a "record check." Researchers in that study compared the people discovered in the census count to those found by drawing individual records (files) at random from sets of administrative records.[56] The resulting studies of coverage and accuracy stretched into the next decade.[57] The results proved useful to those who would rise to challenge the next census.

During the 1960s, the decennial census took on even greater practical significance. Government officials tied census data to algorithms and thresholds that would automatically decide how to allocate power and money. First came the move to automatic apportionment in 1929, then with the New Deal and eventually the Great Society, population figures would channel large flows of federal cash to states and then on to localities. The Nixon and Ford administrations further channeled dwindling streams of resources according to wherever the census said that people could be found.[58] At the same time, Supreme Court decisions (*Baker v. Carr*, *Reynolds v. Sims*, and *Wesberry v. Sanders*, which collectively established "one person, one vote"), followed by the Voting Rights Act of 1965, led to court-enforced numerical tests of discriminatory (and so illegal) gerrymandering in the drawing of congressional districts. According to Anderson and Fienberg, "From the 1940s through the mid-1960s, the [undercount] literature was totally 'methodological'—of interest to demographers, statisticians, and survey researchers," but "this situation was to change dramatically in the mid-1960s."[59] Civil rights activists seized on the undercount to help make their case against ongoing injustices.

A census coordinating committee from Philadelphia, in one instance, used the 1960 undercount measures to push back against claims that the 1970 census was going fine.[60] The committee's chairman detailed all the group had done to try to raise awareness in their community, all their efforts to help get enumerators hired from their communities, their special canvasses of children, and all the pleas they made to the Census Bureau to try new methods to find people, especially if they were Black, or Spanish speakers, in the city. The committee refused to be told that everything was OK, pushing back with the Census Bureau's own studies of the 1960 census: "The experts found that on a national basis the greatest undercount involved adult, black males. ...the undercount was most severe in the larger cities."[61] They demanded a

recount. And they were not alone—"a number of local officials and community groups sued the bureau to enjoin the census and change procedures," while the Bureau registered 1,900 "formal complaints."[62] In contrast, the director of the Census Bureau said, in what would become a familiar refrain: "In my judgment, and in the judgment of the professional staff of the Bureau of the Census, the 1970 decennial census is the most complete census ever taken by this Nation."[63]

As the undercount measurement became a source of controversy, rather than its solution, the Census Bureau sought a way back to comfortable obscurity, a means to mend the undercount's torn veil. In the 1980s, politicians, activists, and statisticians rested new hopes on a plan that would use statistical sampling to not just measure the undercount, but to undo it. The Census Bureau's Undercount Research Staff developed a technique by which individuals counted by a large post-enumeration survey would be matched (or not) to those counted in the ordinary census, such that rates of omission and duplication could be (more or less) precisely calculated and then a suitable number of people added to each area across the nation. This "dual-systems estimation" emerged as a viable possibility in the course of a long-running lawsuit brought by the City and State of New York, which tried to force some sort of adjustment to make up for an undercount in 1980.[64]

But the undercount as a measure still had the power to protect the existing infrastructure and resist change. After the 1990 census, Secretary of Commerce Robert Mosbacher decided against an adjustment. He acknowledged that "Blacks appear to have been undercounted in the 1990 census by 4.8%, Hispanics by 5.2%, Asian-Pacific Islanders by 3.1%, and American Indians by 5.0%, while non-Blacks appear to be undercounted by 1.7%." [65] Yet from those numbers he still decided that the count was "one of the two best censuses ever taken in this country."[66] Mosbacher claimed that it was impossible to statistically adjust the totals without "adversely affecting the integrity of the census." Subsequent challenges to that decision made their way to the Supreme Court where, in a unanimous decision rendered on March 20, 1996, the court affirmed the rejection of adjustment as reasonable.[67]

The undercount measure was generally meant to argue to outsiders: Everything is fine here, carry on. But it also served and serves to justify Census Bureau experimentation and innovation, as officials seek to count more

people more accurately. Congress demanded that the Census Bureau publish its operational plan, justify each decision through research, and document any deviation from the plan, but most technical and statistical procedures only received scrutiny once they became spectacle, and then landed in the Supreme Court. That was the fate of the effort to use administrative records to fill in holes in the enumeration, a practice affirmed by the Court in *Franklin v. Massachusetts* in 1992. That case dealt with the use of military records to count the armed forces temporarily abroad, but it opened the door to use government records more widely. Next came count imputation, which won the Court's approval in 2002 in *Utah v. Evans*.[68] Count imputation has become a particularly important tool for filling in gaps in the count, but without resorting to statistical sampling. It is a procedure that remedies known unknowns. When the Census Bureau is confident that a housing unit is occupied, but cannot get members of that house to respond, they sometimes use imputation to determine the head count and the characteristics of those people. Imputation—and, increasingly, administrative records—allows them to use statistical methods to account for people who don't respond, but only after they make every attempt to reach that household directly. When producing imputed data that are needed for apportionment, every step is taken to ensure that no sampling is involved in the making of these data. The goal instead is to reduce the undercount.

Former Census Bureau directors often talk about who had "the best count ever," which they measure through their understanding of the "the undercount." Yet invoking a single undercount does its own obscuring work. There are, for one, differential undercounts, which focus on comparing the undercounts of demographic subgroups. And there are total undercounts, which measure how far the total population counted differs from the actual population. In a total undercount, the double counting of some people masks the undercounting of others. Large differential undercounts can exist, many people can go uncounted, many people can have been imputed (their race and sex merely guessed), and yet the total undercount can be close to zero, just so long as lots of other folks have been counted twice.

Census data are made, and so are their metrics of success. Pushing the undercount to zero—even if it means turning a blind eye to the double counts—helps reinforce a mirage, an illusion that democracy's data

infrastructure is rock solid. The legitimacy of census data rests, however precariously, on the collectively manufactured notion that the Census Bureau achieves its goal of counting everyone once, and only once, and in the right place. Yet, just as our efforts to "form a more perfect Union" are unending, so too are our efforts to count everyone so that all can be truly represented and receive their fair share of our collective social and financial assets. But once we can all stare into the abyss of how data are made, duct tape will no longer suffice in keeping democracy's data infrastructure whole.

THE GOOSE WHO LAYS THE GOLDEN EGGS OF UNQUESTIONED CONFIDENTIALITY

THIS BRINGS US BACK to the technical disagreement that got us thinking about the census as infrastructure in the first place. Statistical confidentiality is a prime example of a technology intended to allow a complicated infrastructure to fade into the background. The purpose of confidentiality protections is to provide assurances that the data won't be misused, thereby removing any barrier to the smooth collection of personal data.

The modern U.S. Census Bureau's commitment to confidentiality arose in the face of resistance to new kinds of data collection in the early 20th century. The resistance took root among businesses facing new kinds of questions from the Census Bureau in the 1920s. Herbert Hoover, the engineer and rationalizer, sought data (to be provided voluntarily) from individual businesses that could be used to produce statistics about whole industries and about the entire economy, statistics that he hoped would allow for more rational, sustained economic success. Many businesses balked, and Hoover and the bureau tried to woo them with confidentiality.[69]

The rhetorical promise of confidentiality for personal information was older, but it was often breached in the first half of the 20th century. Margo Anderson and Bill Seltzer explain: "Presidents Taft, Wilson, and Hoover all proclaimed statistical confidentiality, and Census Directors in both Republican and Democratic administrations violated the guarantee. … it appears that interagency pressures and perceived national security needs led to weakening the guarantee."[70] Anderson and Seltzer's investigations

have also revealed how the Census Bureau contributed data to assist in the internment of Japanese Americans during World War II under the cover of a War Powers Act that made such actions legal, even as they contravened President Roosevelt's earlier assurances that personal data would not be used to harm individuals.[71]

After the war, fights over access to business data again threatened to undermine the promise of statistical confidentiality. Title 13, first passed in 1954, codified and restated earlier confidentiality protections that prevented the publication or non-statistical use of any individual data. More importantly, the Census Bureau at this time also invested in procedures that would ensure bureaucratic and technical actions aligned with rhetorical and legal commitments. Then the Supreme Court in 1958 sided with the Federal Trade Commission in its efforts to require a paper company to hand over documents it had completed as part of that year's manufacturing census, opening a large bureaucratic loophole in confidentiality protections. Congress closed that loophole in a 1962 revision. Henceforth, the Census Bureau could not publish individual data and no government agency could compel the release of responses prepared for the census.[72] Congress had lent greater legal weight to stand behind the assurance that Richard Nixon would offer in the third paragraph of his 1970 census proclamation:

> Every American can be sure that there will be no improper use of the information given in the Census. Government officials and employees are forbidden by law to use information recorded on the Census form for the purposes of taxation, investigation, regulation, or for any other purpose whatsoever affecting the individual. Every employee of the Census Bureau is prohibited from disclosing information pertaining to any individual.[73]

Yet even as the Census Bureau battled for confidentiality in the courts in the late 1950s, a new set of technical challenges gathered. Academics, and specifically social scientists, threatened the assurance that personal data would be inviolate. In a 1957 document addressed to the American Statistical Association's Census Advisory Committee,[74] census writers cited anonymous data users complaining of the difficulty of running multivariate analyses "by remote control" (that is, by requesting special tabulations from bureau staff

who could access confidential records), while others complained the confidentiality strictures didn't make sense: "It is not clear to me why this duty cannot be performed by suppressing names, addresses, and other possible clues."[75] Bureau officials posited a tradeoff that would allow greater data access: "It may be possible to provide individual cards for the entire country if local-area identification is removed, or to release cards for local areas if personal identification (age, color, occupation, income) is removed."[76]

The desire of academic researchers to get access to fine-grained data manifested itself most prominently in the 1965 proposal for a national data center. In 1965, the Social Science Research Council published a report by the Committee on the Preservation and Use of Economic Data, a report that came to be known as the Ruggles Report, named after the committee's chair, Richard Ruggles. The report heralded a shift in research priorities among economists and other social scientists away from analyzing aggregate data, toward simulation and hypothesis testing using more "basic" individual data.[77] The new social science wanted to build up from individuals, from the micro-economic, and so needed access to micro-data.[78]

Ruggles's committee affirmed the importance of statistical confidentiality and pointed to the Census Bureau for evidence that personal privacy did not have to be the victim of the new social science. The trick was to sacrifice local specificity: "[T]he Census Bureau in the last few years has made available a sample of information on 100,000 individual households, giving considerable detail about the age, education, income ownership, occupation, etc., of the individuals in the household. In this sample, the omission of detailed geographic information makes it impossible to trace the data to any specific individual."[79] This was one solution employed by the bureau to the challenges it faced in the late 1950s.

The price of breaching confidentiality worried bureau officials, though. They did not want to disrupt the appearance of perfect protection. At a 1965 meeting with users of small-area data, Edwin Goldfield, the chief of the Statistical Reports Division within the bureau, warned: "It is constructive to talk about how we can make maximum use of the data while observing the principle of confidentiality but to attempt to solve the problem by eliminating or even infringing upon confidentiality would be to kill the goose that lays the golden egg."[80] Census documents from this period repeatedly explain

that a clear, blanket promise of confidentiality secured more honest answers from individuals and secured them more quickly, and at a lower cost. That's what made confidentiality into Goldfield's goose.

Some of the most exciting potential outcomes that a federal data center could realize also posed distinct threats to confidentiality.[81] Edgar S. Dunn, a consultant hired to advance plans for what was now being called a national data center, noted that better predictive models often depended on bringing together records gathered by different statistical systems, and sometimes required bringing together records about the same individual from different systems.[82] Census statisticians noted that publishing or sharing more data increased the risk that this kind of record linkage might occur more generally, that researchers might discover the identity of individuals by combining information from different sources. A Canadian statistician, I. P. Fellegi, explained in 1970 that "advances in the theory of record linkage, together with the increasing capacity and power of modern computers, represents a new development compelling us to re-evaluate our approaches to disclosure checking in demographic statistics," so that inadvertent or "residual" disclosures of confidential information could be avoided.[83] As Goldfield put it in a letter to his Canadian counterparts: "The doctrine of confidentiality and the procedures for maintaining it keep coming up for review as new needs, new analytical methods and new technology present themselves."[84]

Bureau officials were loath to say too much about how they prevented the identification of individual data. Speaking in 1965, Goldfield talked in broad terms about the ways that the bureau translated broad principles of confidentiality into concrete strategies for "disclosure suppression." They limited geographic specificity in some situations. In others they refused to make certain tabulations that would certainly expose some individual. Then there were other methods that Goldfield could not discuss. "I do not believe," he said, "that it would be desirable for us to attempt to make full disclosure of our disclosure procedures."[85] Disclosure avoidance required disclosure avoidance.

Nothing irks computer security professionals more than this sort of "security by obscurity," where the secrecy of how a system is designed is supposed to ensure that the system is safe from intruders. But that was what disclosure avoidance by data suppression called for, and the Census Bureau

had to extend its use to *all* of the data. Whole tables were suppressed in 1970, with more tables being suppressed in 1980. By 1990, census statisticians added data swapping and "blank and impute" protections, both of which injected arbitrary noise into the census data by either swapping people's records or by erasing someone's record and recreating it through the methods used when people didn't respond. By 2000, they added rounding, top-coding, and other techniques that they did not publicly detail.[86] Each decade, they altered the data a bit more in a race against advances in computing that threatened to unveil tabulated data.

When the Census Bureau started suppressing data, small-area geographies suffered; these were the places where data were too identifiable. Swapping and "blank and impute" strategies created a different problem: bias. In order to protect confidentiality, outliers needed to be harmonized. This meant finding households that didn't look like their neighbors and editing them so that they would. Such an approach was biased, ensuring that communities did not know how diverse they truly were. As communities became less homogeneous—and statistical advances increased—the accepted disclosure avoidance approach threatened to undermine the data's statistical validity.

Disclosure avoidance remained esoteric and mostly went unnoticed, both as an operation and in terms of how it might affect statistics that relied on that data. There weren't decades of public spectacle about how census data were being altered. Instead, slowly, decade-by-decade, the Census Bureau modernized the disclosure avoidance system without anyone paying much attention at all. Those with deep knowledge about the workings of census data understood that the data were being altered, but they also accepted the bureau's arguments that these alterations were not nearly as significant as the error that came from operational challenges, human error, and non-participation. Statistically speaking, this was true, but mostly because the level of error and data editing that takes place before disclosure avoidance is so significant. Even as experts honed their skills to measure undercounts, internal efforts to measure "total survey error" never provided external stakeholders with clear measures of the margin of uncertainty that plagued each and every census record.

While the statistical agency refined its techniques for suppressing data or swapping records, all behind a veil of secrecy, researchers in cryptography

established a new gold standard for security. In their vision of security, truly secure systems remained secure even when publicly interrogated. The software code for a system that relies on public-key encryption, for example, can be published without weakening the system; what protects the system—and its data—is not secrecy, but the mathematical proofs underlying encryption. In the 1990s, computer scientists who had long focused on security started exploring "privacy," by which they meant the technical flow of information. They started developing techniques to undo confidentiality protections while also developing "privacy-enhancing technologies," or PETs, to technically ensure confidentiality, using the same commitment to mathematical proofs that had driven cryptography.

Throughout the U.S. government, scientists and bureaucrats tracked these advances in computing with scientific curiosity and sheer fascination. As law enforcement agencies fretted over how encryption would undermine their intelligence work, statistical agencies recognized that privacy-enhancing technologies might allow them to publicly release previously confidential information without concern that such publications might result in a breach of privacy.

The desires that motivated the national data center plan in the 1960s never faded. Researchers in academia never stopped being frustrated by the lack of access to valuable federal data. To gain access to precious governmental data, they pressured for the public release of more data. When that failed, students and professors often collaborated with government employees, or became federal employees themselves. The Census Bureau's Abowd, a professor of economics at Cornell University, first came to the bureau during sabbatical as a distinguished research fellow in 1998. During his career as a professor, and during his sabbaticals at the bureau, he collaborated with many government economists on a range of topics concerning unemployment and labor force dynamics. Like many of his peers, he kept returning to the bureau because that was where the data was.

Abowd's career had been defined by his struggle to get access to federal economic data; he wanted to make more data publicly accessible for statistical analysis. In 2006, he read a newly published paper by a group of computer scientists who articulated a technique for injecting controlled noise into statistical tabulations that would preserve statistical calculations while also

mathematically guaranteeing privacy.[87] They would call the mathematical proof "differential privacy." Recognizing the potential of this new technique, Abowd gathered his students at Cornell and collaborators at the Census Bureau and hatched a plan. In 2008, his team launched the first significant project that made previously inaccessible Census Bureau data available to the public, protected by differential privacy. The "On the Map" visualization gave researchers unprecedented access to statistical information from the Census Bureau's Longitudinal Employer-Household Dynamics Program.[88]

Census Bureau staff within the Economics and Research & Methodology Directorates were elated by the "On the Map" work. Using differential privacy, they could publish data that had been previously deemed too sensitive for release. Better yet, through a mathematical system, they could manage risk. As scientists, they understood that publishing data "leaked" information, but they saw the key to privacy as being able to trade-off between accuracy and confidentiality. They could statistically account for the risk of reconstruction and reidentification. And still better, they could publish their code, fully assured that technical transparency would not undermine their confidentiality protections. They started to envision all the locked-up data that could now be released for the first time. They also began new collaborations, one of which resulted in the release of data about veterans' post-military career paths.[89]

Meanwhile, other career professionals were exploring the flipside to this data release puzzle, asking how vulnerable existing Census Bureau publications might be to the confidentiality breaches that differential privacy guaranteed against. Shortly before the 2010 census, a group of internal researchers evaluated the protections provided by the disclosure avoidance procedures used for the decennial data releases. They concluded that an attacker could most likely reconstruct many individual records from census data, but it was unlikely that they could link those records to other records to unmask the names of people in the census data. Still, with advances in computing happening rapidly and a flurry of data brokers profiting from data obtained through questionable means, the writing was on the wall: The Census Bureau would need to modernize disclosure avoidance procedures if they wanted to guarantee the same level of data confidentiality that had existed in 2000 and 2010.

In 2015, while preparing for the 2020 census, a group of researchers at the Census Bureau decided to re-evaluate the efficacy of the 2010 disclosure avoidance procedures. What they learned startled them. The team discovered that it was now possible to reconstruct all individual records from the published statistical tables and, more disconcertingly, using only a small fraction of commercial data, to link reconstructed census data with external data to fully reidentify over 50 million people.[90] To their great consternation, they had this level of success using only legitimate commercial data; they feared what they might be able to do using gray-market data. Through this experiment, it became clear to both researchers and senior executives of the Census Bureau that previously used disclosure avoidance methods like swapping were unable to provide a fraction of the protection they once had. Not only were such methods displeasing because they relied on secrecy, and problematic because of their biases, their efficacy for protecting data had been obliterated.

Each decade, the Census Bureau constructs a disclosure avoidance system to protect the confidentiality of individual data as it is processed and turned into statistical tables. Since such systems are never discussed publicly, they never go through the rigorous documentation and testing phases of more publicly visible procedures. For the first time in its history, the Census Bureau decided to invest in designing and developing a robust disclosure avoidance system that could be publicly scrutinized. They decided to start this work early.

Abowd set aside his duties and privileges as a chaired professor in 2016 to go to Suitland, Maryland, and serve full-time as the third chief scientist of the Census Bureau.[91] He was given the mandate of ensuring that the 2020 census data could be kept confidential without harming the range of statistical uses that relied on the accuracy of that data. He did not, at first, recognize the gravity of the task. With "On the Map," he had helped make previously inaccessible data public. This triggered joy and kudos from countless data users, including his peers. But, in order to protect the decennial census data, he was going to have to limit the amount of data that users had come to expect—and add noise to the data that they got. This was not the same task.

The mere fact that the Census Bureau was exploring changes to the disclosure avoidance system that involved differential privacy was never secret,

but it was also never communicated in a way where people understood the radical transformation barreling down the infrastructural tracks. Abowd was focused on the technical puzzle, which was made more complex by constraints that he had not previously encountered. Mathematical proofs do not understand that state counts must be left unaltered. Math does not empathize with policymakers who insist on non-negative integers in population tables even when those cells are filled with very identifiable zeros, ones, and twos. When he spoke publicly, Abowd focused on the technical transformation, highlighting how the confidentiality problem that had long plagued the Census Bureau could finally be solved. But his audience lacked passion for confidentiality. Either they took it for granted or felt that it was an excuse the Census Bureau used to undermine public access to data. Confused by what some dubbed Abowd's "science project," data users had one question on their mind: What would this do to the data?

That question drove Steven Ruggles, son of Richard Ruggles of Ruggles Report fame, to initiate the data user push back against the Census Bureau's new disclosure avoidance plans. As a professor and director of a service to help people access census data known as IPUMS, Ruggles took umbrage at the approach that Abowd and his team were taking. He rallied scholars,[92] orchestrated letters to the Census Bureau leadership that were signed by thousands of scholars, contacted journalists and policymakers,[93] wrote academic articles challenging the method,[94] and lambasted Abowd and the Census Bureau on stage at academic meetings.[95] Like his father, Ruggles has spent his career focused on increasing access to federal statistics for social scientific work. Both father and son felt as though the government had a duty to see its data used for the public good, and both thought strict privacy protections stymied that use. In his presentations, speeches, and comments to journalists, the younger Ruggles argued that, in modernizing its disclosure avoidance procedures, the Census Bureau was going too far, re-interpreting the law to protect more than confidentiality. He argued that they should be stopped.

Ruggles set in motion a backlash against differential privacy, one that rippled from scientists to politicians and advocates. Ever the professor, Abowd responded to the increasingly distressed data users by trying to explain the technical workings of the system, naively believing that if they

understood how the data were being produced, they'd have more confidence in the disclosure avoidance procedures. He spoke lovingly about the beauty of epsilon, the mathematical variable that governs disclosure avoidance. He celebrated transparency, pointing out the ways in which previous disclosure avoidance systems had skewed the data to a degree that data users had never known. None of this reassured anyone. Anxieties escalated as data users tried to interpret what he was saying. Some were horrified by the implication that the data that they had long relied upon had deep flaws; others rejected that premise entirely, content to believe that the previous data had been just fine. And, given that the Census Bureau had never even hinted at the idea that other earlier disclosure avoidance techniques might have introduced biases, the willingness to discuss this now, while touting a new technique, elicited more than a few raised eyebrows.

Abowd doubled down on transparency to reassure data users. He miscalculated. Time and time again, Abowd responded to criticism by sharing more—only what he shared was not what data users wanted to hear. First, the team published the code so that data users could apply it to the formatted 1940 data that had been recently released.[96] Demographers did not know what to make of this offering. Next, Abowd's team applied an early version of the new disclosure avoidance procedures to the 2010 internal census records (pre-disclosure avoidance) and published these as demonstration data so that data users could "see" how disclosure avoidance worked. They were aghast. They found bugs in the system and raised concerns about differences between the published 2010 data and the new demonstration products.[97] What the Census Bureau saw as a good-faith demonstration of a "work-in-progress" was interpreted externally as proof that the Census Bureau's new approach was not fit for use.

Cognizant of the spectacle that had exploded in front of him, Abowd shifted tactics. He began engaging with stakeholders, creating external subcommittees of standing advisory groups, and publishing metrics. He hired a communications person to keep external data users and advocates abreast of what was happening. Yet, having already lost the trust of many data users, much of what Abowd and his team said was viewed suspiciously, as too little and too late. To make matters worse, as partisan interference in the 2020 census increased, the Census Bureau's ability to communicate

anything substantive to anyone became increasingly hampered. Abowd was explicitly muzzled.

As this paper goes into publication in March 2021, the tensions between data users and the Census Bureau continue to be gnarly. Inside the bureau, the technical work of disclosure avoidance improvement is still moving forward, but those civil servants who are responsible for ensuring that test products are of high enough quality must first focus on the underlying data quality problems facing the raw 2020 data. Much remains uncertain as the parameters for the disclosure avoidance system have not been finalized by the senior leaders charged with methodological validity and policy decisions. Consultations, communications, and decisions that were cancelled or postponed during the Trump administration have been reinitiated, but stakeholders are extraordinarily wary. Meanwhile, only days before this paper went to print, the State of Alabama filed suit against the Census Bureau over its disclosure avoidance system.[98] The house is on fire, the firefighters are exhausted, and no one can think about what comes after tomorrow.

Outside the bureau, data users are agitating in different directions. Everyone wants more information, clear operational plans, and reliable timelines. Some wish to seek relief from the courts or from Congress. Others are preparing stakeholder letters to provide the Census Bureau with options that could relieve some of the technical pressure. For many data users, the lack of devotion to addressing their concerns about the disclosure avoidance system—which they believe to be a problem of the Census Bureau's making—is a sign of disrespect and they are agitated. They want the Census Bureau to change its plans, to roll back to previous decades' approaches or come up with a new solution. After months of respecting requests from the civil rights community to be patient as advocates focused on encouraging the public to participate in the census, data users started going to the media and to Congress.[99] Prompted by data users, 33 Democrats from Congress issued a letter to the bureau, demanding that it justify its disclosure avoidance plans.[100] To the media, Congressman A. Donald McEachin of Virginia set in motion a conspiratorial frame, suggesting that differential privacy may have been a political tactic of the Trump administration. Abowd, having been attacked by a Republican member of Congress only a month earlier as a threat to the administration's goals, has the dubious honor of bipartisan hatred.

Speaking to a group of differential privacy experts at a workshop in May 2020, Abowd apologized for his mistakes and his failure to properly understand the broader political context in which he was trying to ensure that confidentiality and statistical analysis could cohabitate.[101] He had thought that finding an acceptable solution to the gnarly problem was possible. He never imagined that "literally no one will be happy." Chastened and exhausted, he wanted this group of experts to know that he had accepted that his reputation would be tarnished through this civil service work. The pain in his voice was palpable. He was, after all, a career civil servant who naively assumed that technical solutions could make a difference.

Nine months later, the technical controversies that formed around those solutions trouble an already troubled operation. The reputation of the Census Bureau has been sullied by the chaos of the pandemic and the spectacle of partisan interventions. The fight over differential privacy threatens to leave it in shambles. The story of this technocratic journey is at its midpoint, defined by strain and anxiety, uncertainty and animosity. What unfolds over the next year could take many different paths, with the legitimacy of the census at stake. Each day brings new statements by the courts, government officials, former census directors, data users, and census advocates. And with each one of these statements, the seeds of doubt about census procedures only grow. Only one thing is for sure: This census will be one for the history books.

CONCLUSION

WE SET OUT TO BETTER UNDERSTAND how a technical controversy might turn into a big deal, one that could shake the infrastructure that supports democratic governance in the United States. Looking at the early 20th century fight over apportionment taught us that political controversy could bring to the surface otherwise obscure technical controversies, transforming them and the seemingly inconsequential scientists involved in them into key actors reshaping a constitutional system. Looking at the late-20th-century battles over the undercount, we saw how even tools intended to discourage critique could land at the center of a firestorm. Throughout it all, we've seen that public data depends on an

infrastructure that often goes unnoticed, systems and processes that are rife with obscure technocratic details so that data can be produced and made "open." Now, as the COVID-19 pandemic and Trump administration meddling has drawn all sorts of scrutiny and skepticism to the census, it seems entirely likely that a different tool intended to discourage critique through increased technological transparency will become a key factor in upcoming debates about American democracy's data infrastructure.

NOTES

1 Julia Lane, *Democratizing Our Data: A Manifesto* (Cambridge, MA: MIT Press, 2020).

2 David Pozen has written about the varied uses of transparency and the way that studies of its efficacy and history make it unwise to consider transparency an end in and of itself, or as a means to necessarily attain good democratic governance. See David Pozen, "Transparency's Ideological Shift," *Yale Law Journal* 128: 100-165, 2018 and Pozen, "Seeing Transparency More Clearly," *Public Administration Review* 80, no. 2 (2019): 326-331.

3 Susan Leigh Star and Karen Ruhleder, "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces," *Information Systems Research* 7, no. 1 (1996): 111-134 at 113.

4 Suzanne Mettler, *The Submerged State: How Invisible Government Policies Undermine American Democracy* (Chicago: University of Chicago Press, 2011) describes partisan investments in rendering government infrastructure invisible. Bureaucrats and stakeholders, seeking to avoid these partisan battles, often attempt to render the infrastructure mundane on purpose.

5 Robert N. Proctor and Londa Schiebinger, *Agnotology: The Making and Unmaking of Ignorance* (Stanford: Stanford University Press, 2008).

6 One of the most famous, and now paradigmatic, controversy studies involved Shapin and Schaffer's study of disputes over the meaning of experiment in early modern England. Steven Shapin and Simon Schaffer, *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (Princeton: Princeton University Press, 1985).

7 The pre-eminent historical account of the twists and turns of the U.S. census across the decades is: Margo Anderson. *The American Census: A Social History* (New Haven: Yale University Press, 1988/2015). There is a substantial literature in history, sociology, media studies, and political science on the messiness underlying seemingly straightforward numbers. See, for instance, Ian Hacking, "Biopower and the Avalanche of Printed Numbers," *Humanities in Society* 5, no. 3 & 4 (1982): 279-295; Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton: Princeton University Press, 1995); Alain Desrosières, *The Politics of Large Numbers: A History of Statistical Reasoning* (Cambridge, MA: Harvard University Press, 2002); Sarah E. Igo, *The Averaged American: Surveys, Citizens, and the Making of a Mass Public* (Cambridge, MA: Harvard University Press, 2007); Wendy Nelson Espeland and Mitchell L. Stevens, "A Sociology of Quantification," *European Journal of Sociology* 49, no. 3 (2008): 401-436; Dan Bouk, *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual* (Chicago: University of Chicago Press, 2015); Jacqueline Wernimont, *Numbered Lives: Life and Death in Quantum Media* (Cambridge, MA: MIT Press, 2018); William Deringer, *Calculated Values: Finance, Politics, and the Quantitative Age* (Cambridge, MA: Harvard University Press, 2018). In terms of the census, much scholarly attention has been devoted to the politics and power dynamics shaping the construction of race within the census. See Melissa Nobles, *Shades of Citizenship: Race and the Census in Modern Politics* (Stanford: Stanford University Press, 2000); Kenneth Prewitt, *What Is Your Race?: The Census and Our Flawed Efforts to Classify Americans* (Princeton: Princeton University Press, 2013); Paul Schor, *Counting Americans: How the US Census Classified the Nation* translated by Lys Ann Weiss (New York: Oxford University Press, 2017); Debra Thompson, *The Schematic State: Race, Transnationalism, and the Politics of the Census* (Cambridge: Cambridge University Press, 2016).

8 The census may not seem like a technical system in the same way an electric grid is, or a nuclear power plant, but the case is easy to make once one considers its complexity and scope, the way both science and law shape it, and its relationship to a variety of technologies ranging from paper forms to maps to tabulating and calculating devices. For a definition of large technical systems, see Thomas P. Hughes, "The Evolution of Large Technological Systems," in *The Social Construction of Technological Systems*, edited by W. E. Bijker, T.P. Hughes, and T. Pinch (Cambridge, MA: MIT Press, 1987): 51-82. For an example of the important place already afforded to the U.S. census in histories of technology, see Martin Campbell-Kelly and William Aspray, *Computer: A*

History of the Information Machine (New York: Basic Books, 1996), chapters 1, 2, and 5. For key works discussing how particular historical circumstances and politics shape technological systems, see Gabrielle Hecht, *The Radiance of France* (Cambridge, MA: MIT Press, 1998); Langdon Winner, "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980): 121-136.

9 Walter F. Willcox, "The Apportionment of Representatives: Annual Address of the President," *American Economic Review* 6, no.1, Supplement (1916): 3-16. https://www.jstor.org/stable/1803789 [https://perma.cc/9M9B-GPBS].

10 Willcox, "The Apportionment of Representatives," 15.

11 Willcox, "The Apportionment of Representatives," 15-16.

12 U.S. House, *Apportionment of Each Number of Representatives from 435 Up to 483, Inclusive by the Method of Major Fractions* document 918, 66th Congress, 3rd Session, 14 December 1920 (Washington, DC: Government Printing Office, 1920).

13 U.S. House, Committee on the Census, *Apportionment of Representatives* 8 January 1921, 66th Congress, 3rd session, report 1173.

14 Huntington to the editors, *New York Times*, 16 January 1921 (Letter was sent/dated on 8 January 1921).

15 Huntington later published the content of this presentation in Edward V. Huntington, "The Mathematical Theory of the Apportionment of Representatives," *Proceedings of the National Academy of Sciences* 7, no. 4 (1921): 123-127.

16 Huntington to the editors, *New York Times*, 16 January 1921.

17 "To Defeat House Increase. Republican Leaders Decide against Reapportionment Bill," *New York Times*, 17 January 1921.

18 *Congressional Record* 60 (1921): 1676-1694 at 1694.

19 Huntington to Howard Sutherland, 22 January 1921, File 412, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

20 Hill to W.S. Rossiter, 14 February 1921, File 408, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC; Advisory Committee to the Director of the Census, "Report upon the Apportionment of Representatives," 18 April 1921 File 408, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

21 The first such bill was discussed in U.S. House Committee on the Census, *Apportionment of Representatives in Congress amongst the Several States* (Washington, DC: GPO, 1926).

22 For Willcox propounding and explaining the scheme in 1927, see Willcox to Dear Sir, 6 June 1927 File 413, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

23 Allyn A. Young to Hill, 6 December 1926 and Allyn A. Young to Hill, 7 January 1927, File 413, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

24 *Congressional Record* 68 (1927): 5414-5430 at 5414-5415.

25 On Vandenberg, see Hendrik Meijer, *Arthur Vandenberg: The Man in the Middle of the American Century* (Chicago: University of Chicago Press, 2017).

26 From Huntington letter as typescript: Huntington, "A Mathematical Error in the Apportionment Bill: The Senate's Opportunity" File 410, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

27 Vandenberg to Huntington, 15 March 1929, File 410, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

28 Vandenberg to Hill, 1 April 1929, File 406, Box 1, Entry 227, Record Group 29, National Archives, Washington, DC; Hill to Huntington, 1 April 1929, File 410, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

29 On the bill's passage, see Meijer, *Arthur Vandenberg*, 57-58.

30 Huntington to Vandenberg, 20 June 1929, File 410, Box 2, Entry 227, Record Group 29, National Archives, Washington, DC.

31 United Press, "Senate Backs House on New Congress Set-Up," *New York Times*, 22 October 1941.

32 Walter F. Willcox, "Area, Population, Birthplace, Migration and Conjugal Condition," *Publications of the American Economic Association* no. 2 (1899): 8-37 at 10-11. http://www.jstor.com/stable/2560482 [https://perma.cc/X6MX-WPPS].

33 Ronald A. Goeken and Diana L. Magnuson, "'I Hate to See the Evening Sun Go Down': Recounts and Padding in the 19th Century U.S. Decennial Census of Population," Social Science History Association Annual Meeting, November 2019 (on file with authors).

34 Irwin Collier, "Harvard. Semester Exams for Statistics. John Cummings, 1896-1900," *Economics in the Rear-View Mirror* http://www.irwincollier.com/harvard-semester-exams-for-statistics-john-cummings-1896-1900/ [https://perma.cc/WYZ8-QKYP].

35 Bureau of the Census, *Negro Population 1790-1915* (Washington, DC: Government Printing Office, 1918) https://hdl.handle.net/2027/uc1.l005114080 4 [https://perma.cc/U68D-7XAT].

36 Bureau of the Census, *Negro Population 1790-1915*, 13. Paul Schor follows the career of Charles E. Hall, noting both his exceptional talents and the opportunities he was denied because of racism. Schor, *Counting Americans*, chapter 18.

37 Bureau of the Census, *Negro Population 1790-1915*, 27.

38 Bureau of the Census, *Negro Population 1790-1915*, 28.

39 Kelly Miller, "Enumeration Errors in Negro Population," *Scientific Monthly* 14, no. 2 (1922): 168-177 at 171-172.

40 Miller, "Enumeration Errors in Negro Population," 173.

41 Miller, "Enumeration Errors in Negro Population," 177.

42 Le Verne Beales, "The Negro Enumeration of 1920: A Reply to Dr. Kelly Miller," *Scientific Monthly* 14, no. 4 (1922): 352-360.

43 Beales, "The Negro Enumeration of 1920," 359.

44 Beales, "The Negro Enumeration of 1920," 360.

45 U.S. House Committee on the Census, *Apportionment of Representatives* (Washington, DC: GPO, 1921), 64-65.

46 U.S. House Committee on the Census, *Apportionment of Representatives in Congress amongst the Several States* (Washington, DC: GPO, 1926), 22.

47 Director Steuart to Allen T. Treadway, 4 February 1929, Folder "Field—Supervisors and Enumerators (Job Descriptions and Qualifications)," Box 211, Entry 210, Record Group 29, National Archives, Washington, DC.

48 On probability and statistics as a 20th century transnational movement, see Theodora Dryer, "Designing Certainty: The Rise of Algorithmic Computing in an Age of Anxiety 1920-1970" (Ph.D. dissertation, University of California, San Diego, 2019).

49 Emmanuel Didier, *America by the Numbers: Quantification, Democracy, and the Birth of National Statistics* Translated by Priya Vari Sen (Cambridge, MA: MIT Press, 2020), 308-319.

50 "Minutes of Census Advisory Committee, 29-30 March 1940," Folder Advisory Committee Meeting March 29 and 30, 1940, Box 76, Entry 148, Record Group 29, National Archives, Washington, DC.

51 Bureau of the Census, *Census of Population: 1950* vol. 1, XIV.

52 Daniel O. Price, "A Check on Underenumeration in the 1940 Census," *American Sociological Review* 12, no. 1 (1947): 44-49.

53 Price, "A Check on Underenumeration in the 1940 Census," 49.

54 Prewitt, *What is Your Race?*, 106-113.

55 Frederick F. Stephan to Morris H. Hansen, 2 August 1957, page 6, Folder 5, Box 3, Entry A1-389-E, Record Group 29, National Archives, Washington, DC.

56 U.S. Bureau of the Census, *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Record Check Studies of Population Coverage* (Washington, DC: GPO, 1964) https://hdl.handle.net/2027/mdp.39015055377447 [https://perma.cc/L6WX-BPAJ]; for an overview of the 1960 accuracy research program, see U.S. Bureau of the Census, *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Background, Procedures, and Forms* (Washington, DC: GPO, 1964) https://hdl.handle.net/2027/mdp.39015055377462?urlappend=%3Bseq=7 [https://perma.cc/5VLR-SLV6].

57 These 1960 analyses were still being published as late as 1973, including this one originally led by Barbara Bailar: U.S. Bureau of the Census, *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics for Subgroups of the Total*

Population (Washington, DC: GPO, 1973) https://hdl.handle.net/2027/uc1.d0003243706?urlappend=%3Bseq=5 [https://perma.cc/SAS5-UCLA].

58 Keeanga-Yamahtta Taylor, *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership* (Chapel Hill: University of North Carolina Press, 2019), 244-248.

59 Margo J. Anderson and Stephen E. Fienberg, *Who Counts?: The Politics of Census-Taking in Contemporary America* (New York: Russell Sage Foundation, 1999), 31.

60 U.S. House Committee on Post Office and Civil Service, *Accuracy of 1970 Census Enumeration and Related Matters* 1970, no. 91-30, 451-454.

61 U.S. House Committee, *Accuracy of 1970 Census Enumeration and Related Matters*, 453.

62 Anderson and Fienberg, *Who Counts?*, 40.

63 U.S. House Committee, *Accuracy of 1970 Census Enumeration and Related Matters*, 3.

64 On dual-systems and the court case, see esp. Anderson & Fienberg, *Who Counts?*, 50, 79-80, 88.

65 Federal Register 56, no. 140 (1991): 33582.

66 Federal Register 56, no. 140 (1991): 33582.

67 Anderson and Fienberg, *Who Counts?*, 163-164.

68 Patrick J. Cantwell, Howard Hogan, and Kathleen M. Styles, "The Use of Statistical Methods in the U.S. Census: 'Utah v. Evans'" *American Statistician* 58, no. 3 (2004): 203-212.

69 See, for instance, the resistance of woolens manufacturers to Census Bureau requests in Folder 31-hhcom-b126-f2207 "Commerce Department Census, Bureau of 1924 January-April," Herbert Hoover Presidential Library, West Branch, IA.

70 Margo Anderson and William Seltzer, "Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965," *Journal of Official Statistics* 23, no. 1 (2007): 1-34 at 10.

71 Anderson and Seltzer, "Challenges to the Confidentiality of U.S. Federal Statistics," 14-23; William Seltzer and Margo Anderson, "Census Confidentiality under the Second War Powers Act (1942-1947)," Paper prepared for presentation at the session on "Confidentiality, Privacy, and Ethical Issues in Demographic Data," Population Association of America Annual Meeting, March 29-31, 2007, New York, NY.

72 Anderson and Seltzer, "Challenges to the Confidentiality of U.S. Federal Statistics," 25-27.

73 Richard Nixon, Proclamation 3973, 1970.

74 U.S. Census Bureau, "Statement on Problems Connected with Modifications of Statutes or Procedures Relating to Disclosure of Census Data," 4 January 1957, Folder "Confidentiality Materials," Box 10, Entry A1 356, Record Group 29, National Archives, Washington, DC.

75 U.S. Census Bureau, "Statement on Problems," 5.

76 U.S. Census Bureau, "Statement on Problems," 11.

77 While statistics traditionally dealt in aggregated wholes, this new wave wanted to work (to the extent possible) with information on unaggregated individuals, building a record of the nation with new and vast collections of "data doubles," those pieces of personal data that remain tied to an individual or household, the kind of data that statistical confidentiality protected. See Sarah E. Igo, *Known Citizen: A History of Privacy in Modern America* (Cambridge, MA: Harvard University Press, 2018), chapter 6; Dan Bouk, "The History and Political Economy of Personal Data over the Last Two Centuries in Three Acts," *Osiris* 32 (2017): 85-106; Dan Bouk, "The National Data Center and the Rise of the Data Double," *HSNS* 48, no. 5 (2018): 627-636.

78 The Ruggles Report was published as Social Science Research Council, *Report of the Committee on the Preservation and Use of Economic Data*, April 1965, https://ia800200.us.archive.org/31/items/ReportOfTheCommitteeOnThePreservationAndUseOfEconomicData1965/Ruggles_econdata_1965.pdf [https://perma.cc/AAM6-RSQW].

79 Social Science Research Council, *Report*, 17-18.

80 U.S. Census Bureau, "Bureau of the Census Administrative Manual: Confidential Nature of Data Collected by the Bureau of the Census," 19 June 1968, appendix C, Folder "Confidentiality Materials" Box 10 Entry A1 356, Record Group 29, National Archives, Washington, DC.

81 In *The Known Citizen*, chapter 6, historian Sarah Igo details how the proposal for a national data center helped trigger a backlash, which led to Congress passing a series of privacy-related laws and reforms in the 1970s. None of these new laws altered Title 13,

which provides even greater privacy protections for census data.

82 Edgar S. Dunn Jr., and Office of Statistical Standards, Bureau of the Budget, "Review of Proposal for a National Data Center," *Statistical Evaluation Report* 6 (Dec 1965), 9-10. https://hdl.handle.net/2027/coo.31924013916055?urlappend=%3Bseq=88 [https://perma.cc/442Q-W4YC].

83 I. P. Fellegi from the Dominion Bureau of Statistics, "On the Questions of Statistical Confidentiality" 14 September 1970, page 8, Folder "Confidentiality Materials," Box 10 Entry A1 356, Record Group 29, National Archives, Washington, DC.

84 Edwin D. Goldfield to Lorne Rowebottom (asst. Dominion Statistican in Ottawa), 10 February 1965, page 2, Folder "Confidentiality Materials" Box 10 Entry A1 356, Record Group 29, National Archives, Washington, DC.

85 U.S. Census Bureau, "Bureau of the Census Administrative Manual: Confidential Nature of Data Collected by the Bureau of the Census," 19 June 1968, appendix C, page 3, Folder "Confidentiality Materials," Box 10, Entry A1 356, Record Group 29, National Archives, Washington, DC.

86 U.S. Census Bureau, "A History of Census Privacy Protections," 10 October 2019, https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html [https://perma.cc/B4KA-Q4W9].

87 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, "Calibrating noise to sensitivity in private data analysis," in Proceedings of the Third Conference on Theory of Cryptography (TCC'06), edited by Shai Halevi and Tal Rabin (Berlin: Springer-Verlag, 2006): 265–284.

88 See: John M. Abowd, "Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau," 17 August 2018, https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html [https://perma.cc/F3NX-LFD3]; and U.S. Census Bureau, On the Map, https://onthemap.ces.census.gov/ [https://perma.cc/5YKT-HBU5].

89 See: U.S. Census Bureau, *Veteran Employment Outcomes Explorer*, https://lehd.ces.census.gov/applications/veo/occupation/detail/ [https://perma.cc/PF3U-ZUP3].

90 Some of the reconstructed records did not match the original data collected because of previous disclosure avoidance protections; this prevented a larger number of people from being reidentified. Furthermore, because they only worked with a small fraction of data available in the wild, an attacker with richer data could've reidentified many more people. See: Simson Garfinkel, John Abowd, and Christian Martindale, "Understanding Database Reconstruction Attacks on Public Data," *ACMQueue* 16, no. 5 (2018): 28-53.

91 The Research and Methodology Directorate of the Census Bureau was reconstituted in 2011 after the methodological work being done attracted political ire. The Chief Economist in the Economic Directorate–Ron Jarmin–was put in charge and tasked with advancing the methodological rigor of the work done at the Census Bureau. To integrate scientific advancements from academia, Jarmin began inviting scholars to serve as Chief Scientists on a rolling basis.

92 An article in Science was one of the many written to communicate about differential privacy with scholars: https://science.sciencemag.org/content/363/6423/114.full [https://perma.cc/TJ8M-2R3B]. Ruggles also regularly discussed this topic on his mailing list to IPUMS users, which can be subscribed to through https://usa.ipums.org/usa/index.shtml [https://perma.cc/2MZK-7DTB].

93 For example: https://www.sciencemag.org/news/2019/01/can-set-equations-keep-us-census-data-private [https://perma.cc/N5MU-D69H].

94 Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder,"Differential Privacy and Census Data: Implications for Social and Economic Research," *AEA Papers and Proceedings* 109 (2019): 403-408.

95 One such session was at *Population Association of America* in Austin, TX in April 2019.

96 In 1978, Congress passed a law that requires Census Bureau data to be stored at the National Archives for 72 years. After that, the data is to be released to the public. See: U.S. Census Bureau, "The '72-Year Rule'," https://www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html [https://perma.cc/EH4L-QN4J].

97 National Academies of Sciences, Engineering, and Medicine, "2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop" (Washington, DC: The National Academies Press) https://doi.org/10.17226/25978 [https://perma.cc/E83U-HCKR].

98 See: "Court Case Tracker: Alabama v. U.S. Dep't of Commerce," Brennan Center for Justice, last modified March 15, 2021, https://www.brennancenter.org/our-work/court-cases/alabama-v-us-dept-commerce [https://perma.cc/65FS-JD8E].

99 Tierney Sneed, "The Census 'Time Bomb' Awaiting Joe Biden," *TPM*, 16 December 2020 https://talkingpointsmemo.com/news/census-differential-privacy-criticisms [https://perma.cc/PE6W-Q8CQ].

100 Mel Leonor, "McEachin Casts Doubt on Census Bureau's Plans for Population Count Data," *Richmond Times-Dispatch*, 21 September 2020 https://richmond.com/news/state-and-regional/govt-and-politics/mceachin-casts-doubt-on-census-bureaus-plans-for-population-count-data/article_78b9dd06-c21d-5acc-b564-00fbc5934607.html [https://perma.cc/G7MG-97EH].

101 boyd attended the session on 12 May 2020 in which Abowd apologized.

About the Authors

Dan Bouk is Associate Professor of History at Colgate University and a Faculty Fellow at Data & Society. He researches the history of bureaucracies, quantification, and other modern things shrouded in cloaks of boringness. His work investigates the ways that corporations, states, and the experts they employ have used, abused, made, and re-made the categories that structure our daily experiences of being human. His first book, *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual* (University of Chicago Press, 2015), explored the spread into ordinary Americans' lives of the U.S. life insurance industry's methods for quantifying people, discriminating by race, justifying inequality, and thinking statistically. His recent writings put today's political and economic values of personal data in a much wider historical context. He is currently completing a book on the U.S. census for MCDxFSG titled *Democracy's Data and How to Read It*.

danah boyd is a Partner Researcher at Microsoft Research, the founder and president of Data & Society, and a Visiting Professor at New York University. Her research is focused on addressing social and cultural inequities by understanding the relationship between technology and society. Her most recent book—*It's Complicated: The Social Lives of Networked Teens and Participatory Culture in a Networked Age* (Yale University Press, 2014)—examines the intersection of everyday practices and social media. She is a 2011 Young Global Leader of the World Economic Forum, a member of the Council on Foreign Relations, a Director of both Crisis Text Line and Social Science Research Council, and a Trustee of the National Museum of the American Indian. She received a bachelor's degree in computer science from Brown University, a master's degree from the MIT Media Lab, and a Ph.D. in Information from the University of California, Berkeley.

© 2021, Dan Bouk and danah boyd.

Acknowledgments

This paper is based on original research and fieldwork by the authors, and would not have been possible without the support of many people. First, we want to acknowledge the civil servants at the Census Bureau and a range of advocates and stakeholders who helped us understand different dimensions of the stories that we tell. We presented earlier drafts at three workshops that helped us strengthen our arguments: Privacy Law Scholars Conference, Society for the Social Studies of Science (4S), and Data & Democracy. The discussants, commentators, and co-panelists at these events provided invaluable feedback. Katy Glenn Bass and Amy Kapczynski provided crucial feedback and editorial advice. Finally, our team at Data & Society have provided invaluable support throughout this process. This paper was made possible through funding provided by the Alfred P. Sloan Foundation and the John S. and James L. Knight Foundation.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Illustration: ©Erik Carter



KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY