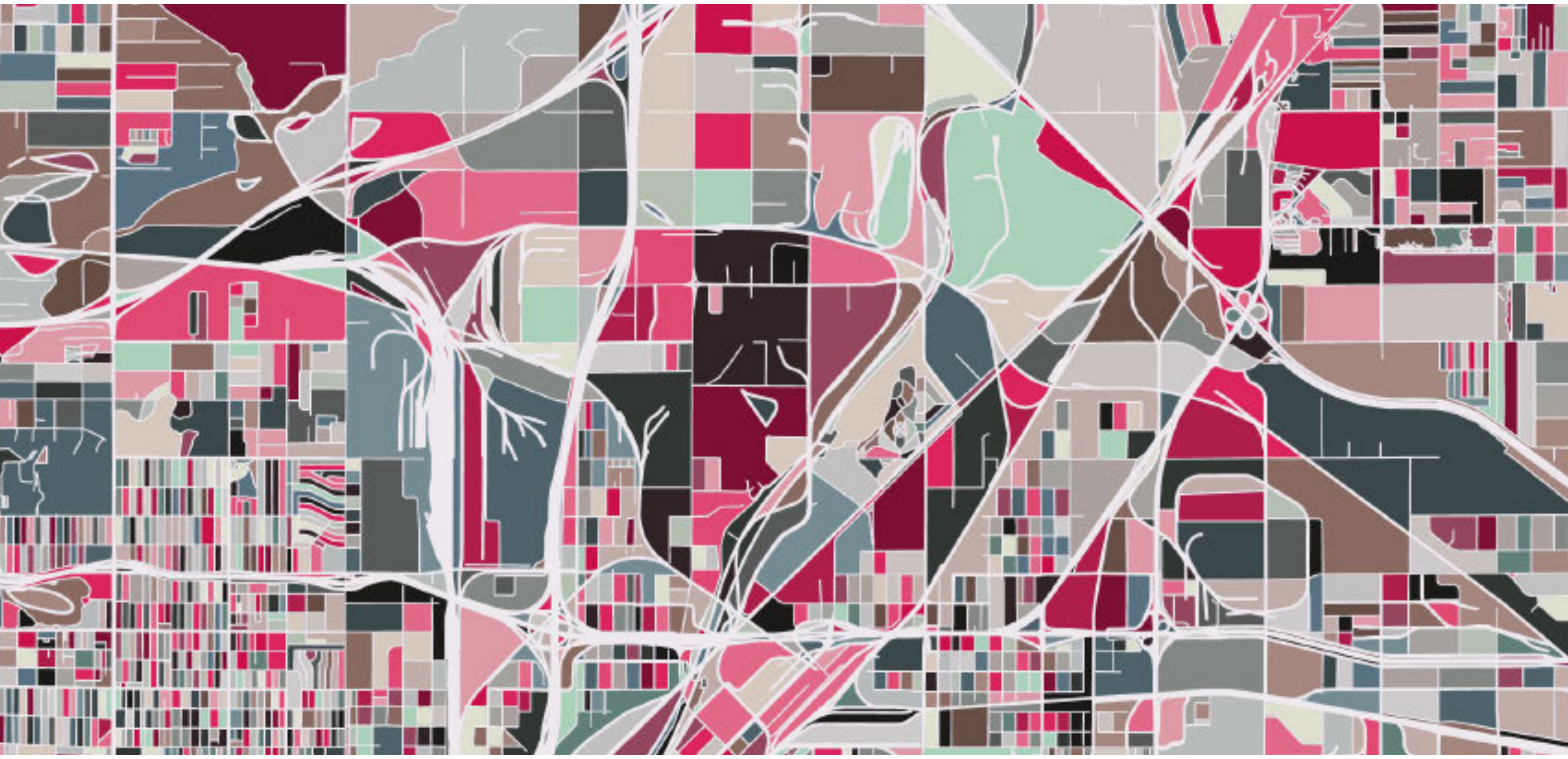
DATA&
SOCIETY

# Balancing Data Utility and Confidentiality in the 2020 US Census

**A LIVING DOCUMENT BY**
danah boyd



This paper is crafted as a living document, capturing how confidentiality and disclosure avoidance are managed in the 2020 US Census. It records an evolving set of responses and actions over time, and will be updated as new information becomes public.

**THIS VERSION LAST UPDATED ON DECEMBER 10, 2019.**

DATA&
SOCIETY

**Table of Contents**

Author: danah boyd, founder and president, Data and Society, and partner researcher, Microsoft Research; PhD, 2008, School of Information, University of California at Berkeley.

*This paper is crafted as a living document, capturing how confidentiality and disclosure avoidance are managed in the 2020 US Census. It records an evolving set of responses and actions over time, and will be updated as new information becomes public. This version last updated on December 10, 2019.*

This research is supported by the Alfred P. Sloan Foundation, as well as funders of Data & Society; for more information on Data & Society's funders, please visit https://datasociety.net/about/#funding.

## Introduction

The United States census is conducted every decade to support the apportionment of the House of Representatives, the allocation of federal tax dollars, and the redistricting within each state. The data that are collected during this process are released as statistical tables, which also support numerous other activities, including social science research and local policy-making. Over the last century, these census data products have become more and more refined, increasingly publicly accessible, and even more widely used.[1]

As the Census Bureau prepares to enumerate the population of the United States in 2020, the bureau's leadership has announced that there will make significant changes to the statistical tables the bureau intends to publish. Because of advances in computer science and the widespread availability of commercial data, the techniques that the bureau has historically used to protect the confidentiality of individual data points can no longer withstand new approaches for reconstructing and reidentifying confidential data. As will be discussed in more detail, research at the Census Bureau has shown that it is now possible to reconstruct information about and reidentify a sizeable number of people from publicly available statistical tables. The old data privacy protections simply don't work anymore. As such, Census Bureau leadership has accepted that they cannot continue with their current approach and wait until 2030 to make changes; they have decided to invest in a new approach to guaranteeing privacy that will significantly transform how the Census Bureau produces statistics.

Described as "formal privacy" or "differential privacy," the new disclosure avoidance system being built by the Census Bureau allows the disclosure avoidance team to mathematically assess the risk that released statistical tables could be used to reidentify individuals. This allows the bureau to rigorously and transparently guarantee a qualified level of privacy protection in ways that are not possible using traditional disclosure avoidance methods. While this transformation of census data products should be reassuring because it *guarantees* a defined measure of confidentiality, many data users are unsure of what these changes will mean for the products they regularly use for funding, redistricting, policy making, and social science. Moreover, given the technical details involved, many census stakeholders simply do not know what to think. And given the importance

[1] At the International Conference on Machine Learning on June 20, 2019, John Abowd stated that the Census Bureau released over 150 billion statistics after the 2010 census. The relevant slide can be found at 8:20 in the YouTube version of his talk: https://www.youtube.com/watch?v=R_8riuhlw-4.

of the bureau's statistical tables to democracy, resource allocation, justice, and research, confusion about what differential privacy is and how it might alter or eliminate data products could inadvertently undermine trust in the Census Bureau and its processes. For this reason, it is important to better understand what is unfolding and why.

This paper is a living document, designed to provide context to the Census Bureau's decision to integrate differential privacy into its processes for producing data products. This decision has been controversial, in part because of how it was made and communicated. From the bureau's perspective, the Bureau's decision is borne out of its deep and statutorily mandated commitment to protecting the confidentiality of individual data records in order to maintain the trust of the public, whose cooperation and data are needed for the census to be successful. The decision to proceed with implementing differential privacy now, even without everything finalized, is driven by significant transformations in computational power and commercial data that make decennial data more vulnerable than in previous censuses.

Still, this decision introduces significant risks. First, the technical implementation is not finished. To date, employees of the Census Bureau have regularly offered statements like "we don't yet know" to important operations questions, which increases anxiety among data users and other census stakeholders. Second, for many stakeholders, the idea that data utility and privacy can be operationalized into a mathematical trade-off is both unfathomable and deeply unsettling. Third, because there is no way to guarantee confidentiality and still produce all of the tables published in the past, the Census Bureau has stated unequivocally that there will be more limits on what data is produced than there were previously. This seeds antagonism between different stakeholders who are all passionate about the census. Fourth, uncertainty about what data products will be produced – and what they will look like – is nerve-racking for data users whose work relies on the Census Bureau delivering data in a particular format on a particular date. Changes have material ramifications for data users who have to operate on a very specific timeframe for redistricting and allocating funds. Finally, the approach the Census Bureau is taking allows for maximal transparency about process, which inadvertently draws attention to some of its earlier data privacy work. While a move toward greater transparency should be celebrated, it also risks creating new forms of doubt.

These issues are all deeply entwined and reveal how difficult it is to govern a system of data infrastructure like the census. The goal of this paper is to document and clarify the different issues at play and shed light on what is known about how the Census Bureau is trying to balance data utility with confidentiality as the 2020 decennial census approaches. Because much is not yet known and because the issues described in this document are evolving, this paper is intended to be a living document that will evolve as the system matures and the debates crystalize. Much work is needed to bridge discussions between the different stakeholders who are invested in the trade-off between confidentiality and data utility; this paper is intended to help ground those discussions.

### Background: Census, Confidentiality, and Trust

Every 10 years, the United States produces a census of all people living in the country.[2] The census, which has been conducted every decade since 1790, is essential to representative apportionment, redistricting, and the fair distribution of shared financial resources. Participation in the census is required by law to uphold the constitutional mandate of counting everyone. There have been many technical changes to the census over the decades, often to increase the accuracy and availability of data or to better protect individual privacy. The census was once conducted solely through door-to-door enumeration, but since 1970, there has been a self-response process in the form of a paper questionnaire returned by mail.[3] The 2020 Census will be the first decennial census conducted with four potential means of core participation: internet, telephone, and paper self-response, and door-to-door enumeration.[4] The last, known now as nonresponse follow-up (NRFU), is conducted after the self-response phase to make certain that every resident (defined as a person usually sleeping in a "living quarters address" in the "master address file") is counted.

The innovation of paper-based self-response emerged in the mid-20th century for multiple reasons. The data were believed to be more accurate when people filled out the forms in the privacy of their own homes rather than sharing information with a stranger at the door. This "modern" way of collecting data also increased convenience (and, perhaps, trust). It was also believed to lower the cost of conducting the census compared to universal field enumeration.

[2] For a detailed historical account of the census, see: Anderson, Margo J. 2015. *The American Census: A Social History, Second Edition*. Yale University Press.

[3] In 1960, the paper census was sent out by mail but enumerators went door to door to collect the paper forms.

[4] There are additional operations for specific subpopulations, such as those living in group quarters and the unhoused.

The census – also known as the decennial census – is not the only data collection effort that the Census Bureau conducts, but it is unquestionably the most expensive, complex, and politically contentious operation. Moreover, it is no longer the only attempt by the Census Bureau to understand household demographics. From 1790 until 1930, a period when the decennial census was the only household data collection effort, increasingly more questions were added to support the work done by policy makers, businesses, and social scientists. There were questions about income, employment, plumbing, literacy, mental health, and radio ownership. In 1940, the bureau started producing two forms—a short form and the a longer one. The former was asked of everyone and focused on a small number of demographic items.  The latter included extensive questions but was completed only by a sample of the total population. This was possible thanks to advances in statistics. Starting with the 2010 census, the basic data are still collected via the decennial operation, but the responsibility for all other data collection has shifted to the American Community Survey (ACS), a survey that samples a portion of the population and takes place on a rolling basis throughout the decade. While the apportionment of representatives requires using only decennial data, most other federal laws and data users can – and often do – use data from the ACS, as well as annually updated population estimates built off the decennial census.

Trust has always been important to the bureau. After all, the data released must be trusted by the public and by Congress. When the census becomes misused or politicized, trust in the process weakens.[5] Although collecting data for the census was once a very public activity, by the 20th century, the bureau started taking numerous steps to ensure confidentiality of the data and the data collection process.[6] One important step occurred in 1954, when Congress enacted Title 13 of the US Code.[7] Title 13 sets forth details about the administration of the census, confidentiality of census data, and penalties for violating confidentiality.[8] Of particular note, 13 USC §9 states that "Neither the Secretary, nor any other officer or employee of the Department of Commerce or

[5] Throughout the 20th century, there were a number of incidents that undermined trust in the Census Bureau and the data themselves. One example concerned the 1940 Census, where data were provided to government officials to help locate people who were deemed to be a threat to national sovereignty during World War II, namely Japanese Americans. The human rights atrocities that ensued were not connected back to the census until decades later. Although today's laws prohibit a similar violation of confidentiality, those who have witnessed governmental violations of individual rights are often wary of the strength of legal protections, and they point to this case as an example. For more details on the relationship between the Census Bureau and Japanese internment, see Christa Jones' 2017 memo "Original Sources and Research Concerning Census Bureau Efforts to Support Japanese Internment": https://ecommons.cornell.edu/handle/1813/66186.

[6] In *The Known Citizen,* Sarah Igo offers a historic accounting of privacy in light of government data collection. Igo, Sarah E. 2018. *The Known Citizen: A History of Privacy in Modern America*. Cambridge, Massachusetts: Harvard University Press, 2018.

[7] Title 13 is available at: https://www.govinfo.gov/content/pkg/USCODE-2007-title13/pdf/USCODE-2007-title13.pdf.

[8] In effect, Title 13 formalized the confidentiality norms that were already in place at the Census Bureau. For a more detailed history, see Anderson, Margo and William Seltzer. 2007. "Challenges to the Confidentiality of US Federal Statistics, 1910-1965." *Journal of Official Statistics* 23(1), 1-34.

bureau or agency…may… make any publication whereby the data furnished by any particular establishment or individual under this title can be identified.” In short, the bureau must produce statistics for the nation, but personal information cannot be given to government agencies, law enforcement personnel, immigration services, or courts of law. Per Title 13, all who work for the bureau are sworn to uphold the confidentiality of the data (per Title 13) for life and are subject to strict penalties, including fines and imprisonment. This oath is taken very seriously in the halls of the Census Bureau.

The bureau releases aggregated and anonymized data products, but does not immediately make the underlying data available to the public or other governmental agencies.  Individuals named on a record (or their heirs) can examine original records upon request. Otherwise, personally identifiable census information remains confidential for 72 years, after which the National Archives and Records Administration releases all individual records of a given decennial to the public. The most recent census currently available for public inspection is the 1940 Census.
In the meantime, no non-Census Bureau employee – including employees of other governmental agencies, academic researchers, or member of the public – can access confidential information. Moreover, Census Bureau employees can only access confidential data if they propose and are approved to do a project that is deemed beneficial for the statistical work of the agency.[9]

While confidentiality is managed by both legal and procedural protections, as well as by bureaucratic and technical ones, a public’s trust in an institution like the Census Bureau is not wholly determined by the steps it takes to ensure confidentiality. Fundamentally, the political climate during any given census shapes public perception, which in turn affects both the processes for conducting the census and the attitudes stakeholders have about the quality of the census data. Although the Census Bureau is comprised almost exclusively of nonpartisan professionals from a range of scientific disciplines, the bureau reports to the Department of Commerce, which is an agency of the executive branch. Depending on who is president – and how the public feels about that administration – the public may perceive Census Bureau actions in a particular light. This has been true throughout the Census Bureau’s history.

[9] All such employees are obliged to uphold confidentiality per Title 13; they must also receive approval for the public dissemination of their results.

## A Complex Operation: From Data Collection to Data Products

The activities that are undertaken to produce census data products are complex; the Census Bureau regularly touts the notion that the census is the largest peacetime operation that the US does. While the full details of the census operations are out of scope for this paper[10], this section will provide a brief overview of how the data are created and protected.

A wide range of operations go into data collection. Most people first recognize that the census is underway when a range of stakeholders make a concerted effort to "get out the count" (GOTC) in advance of the "self-response" phase. In particular, there will be numerous events, media campaigns, and news stories leading up to Census Day on April 1 (which is also referred to as "reference day"). The bulk of people will participate, either through self-response or NRFU.
The 2020 self-response push will encourages people to participate online, by phone, or by mailing in a paper form. Self-response outreach will begin in mid-March 2020 and continue on through May of that year.  After assessing which households do not respond based on their "Master Address File (MAF)," the bureau will then begin to send census workers out to specific houses to collect data as part of the NRFU operation. Should people still fail to respond, the bureau will turn to a combination of proxies (e.g., neighbors) and administrative government records to try to fill in the gaps.

While this self-response/NRFU household count is the core operation, a range of other operations also take place. For example, there is a special operation to count remote Alaska that begins in January, a "group quarters (GQ)" operation that focuses on people living in shared residential housing, a program to count military living oversea, and a two-day effort to count all who are unhoused. All of this data is collected in order to produce the Decennial Response File (DRF).

Most of the census operations focus on collecting data about *households,* not individuals per say. For each household, one person is expected to respond (Person 1). They are asked to indicate if the living quarters is an apartment, home, or mobile home—and if it is rented, owned, or mortgaged. Then every person living in the housing unit is to be listed, along with a set of characteristics. The census collects name, sex, date of birth, Hispanic origin, and race. For all but

[10] For a detailed accounting of the census operational plan, see https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/operational-plan.html.

Person 1, the census also asks if the person typically lives elsewhere and how this person is related to Person 1.

Each household is assigned to an address that is listed in the MAF. When new addresses are found during one of the operations, they are added to the MAF. An address is marked "delete" on the MAF if the address no longer exists and "vacant" if a census worker approaches the address and no one is living there. At the end of the data collection phase, the bureau examines the MAF to see which addresses are unresolved, meaning that there are either with *multiple* responses for a single address or marked as occupied but with no known people. The bureau must resolve all addresses on the MAF to proceed, even if it means drawing on administrative records, asking a proxy neighbor for an estimate of how many people dwell there, determining which duplicate record is more likely to be accurate, or imputing the household count based on the statistics in the same geography for the same kind of dwelling.

After resolving every address in the MAF, the bureau produces the "Census Unedited File (CUF)." It is the CUF that is used to calculate congressional apportionment. Once a state count is locked in the CUF, the Census Bureau prepares the first data product—the state population counts. The number the Census Bureau produces is based purely on the census count (resident population plus those living with federal employees stationed abroad) and will be delivered to the president no later than December 31, 2020. This data is then delivered to Congress. As dictated in the 1929 Reapportionment Act, Congress has 15 days to either apply the Hill-Huntington method[11] to determine how many of the 435 seats in the House of Representatives goes to each state and territory. Alternatively, they could pass a law to change the 1929 Reapportionment Act. The Census Bureau does the Hill-Huntington calculation as a courtesy. Then, the Secretary of the House must notify each state of its apportionment.

Meanwhile, the Census Bureau begins its work to produce other data products. The next two data products that are to be released – on March 31, 2021 – are known as the PL 94-171 and the Citizen Voting Age Population (CVAP).[12] The former is named after Public Law 94-171 and consists of the data necessary for redistricting. By custom, and at the request of specific states and the Secretary of Commerce, the Census Bureau, the Census Bureau prepares data down to the geographic block

[11] U.S. Census Bureau, "Methods of Apportionment," https://www.census.gov/history/www/reference/apportionment/methods_of_apportionment.html.

[12] In 2010, the CVAP file was produced by the American Community Survey Directorate, using those data about citizenship and voting age (18+); that data is only provided at the block-group level. How the CVAP will be produced for 2020 is still unknown. As of 2019, the Secretary of Commerce has directed the Census Bureau to produce this file down to the block level which requires a significant transformation of its processes.

level. This data includes critical characteristics of each person in each household because this data product must contain both housing unit and population counts down to the geographic block level, as well as tabulated racial and Hispanic-origin characteristics of people in the required geographies.

Census geography is built on a hierarchy (see Figure 1). The smallest unit is the block. Every other geography unit builds from the block. In urban settings, a census block is often literally a city block – the set of residences clustered between streets – but in other locations, blocks can stretch to accommodate larger areas. In 2010, there were over 11 million census blocks (but almost 5 million of them contained zero people). Census blocks are combined to form block groups. Block groups, in turn, are combined to form census tracts, which are combined to form census counties, which are combined into states. These census geographies fit together to produce the census spine.[13]

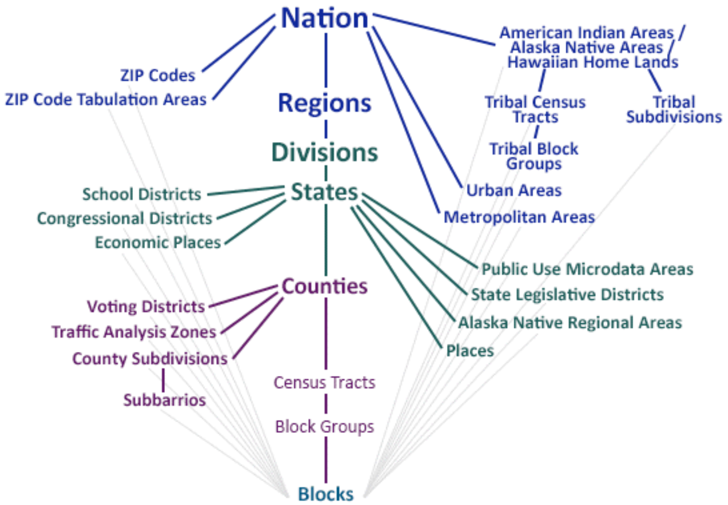


*Figure 1 – Census geographies. In the middle is the census spine,” or the standard hierarchy of census geography entities from the smallest (block) to the largest (nation). Other geographies are considered off the spine” even though they build from census blocks and connect into the spine at higher levels of geography. Source: U.S. Census Bureau, Oct. 27, 2010.*

[13] Consistent national-level census geographies emerged in the 20th century and have shifted over time. The geographies supplied here are the ones used in the present. Census counties often functionally equivalent to the political counties defined by the states, but not always; Alaskan census areas are a notable exception. For a brief history of the development of census geographies, see US Census Bureau. "Tracts and Block Numbering Areas." https://www.census.gov/history/www/programs/geography/tracts_and_block_numbering_areas.html.

There are a range of other geographies that are off the spine, including school districts, tribal lands, cities, and voting districts, etc. Most of these geographies are pulled back into the spine at the county or state level. For example, congressional districts are comprised of blocks, and may be combined from partial block-groups, but no Congressional district cuts across a state. The American Indian and Alaska Native geographies are unique because they can – and do – cross states. These geographies are still comprised of census blocks. All individuals – and all households – are assigned to a single census block, but only connect back to the spine at the national level. All addresses on the MAF are assigned to a single census block. All individual counts at each address on each block are resolved before the production of the CUF.

After producing the CUF, the Census Bureau must resolve missing and inaccurate demographic data before producing the "Census Edited File (CEF)," which is what is used to produce the other data products, including the PL 94-171 redistricting file. To produce the CEF, every cell for every person and every household must have a value. To achieve this, the bureau must resolve conflicting data and fill in missing data. The cleanup process to produce the CEF takes multiple forms. Consider people who indicate that their date of birth does not result in the age they list. That conflict can sometimes be addressed through administrative records, such as birth records and social security records. Sometimes, people put a child as Person 1, such that a 35-year-old is the "child" of a 3-year-old. That is resolved by changing who is designated as Person 1. Sometimes, during self-response, people check all of the boxes. This may be ignored, or it can be resolved by using responses from the same person to previous census questionnaires or administrative records. Still, after all of this work, some data are still missing. At that point, the Census begins assigning and imputing data. Assignment involves taking data from earlier census responses or administrative records from the same household. The details of imputation are not publicly available, but, it appears as though the bureau builds statistical models to provide "best guess" answers to missing fields, based on signals from others in the same household, block, and broader geography. There are many different types of imputation that the bureau uses to produce the CEF, all informed by the quality standards enforced by the Office of Management and Budget (OMB) through the Chief Statistician of the United States.[14] The bureau has also conducted extensive research to inform its imputation process and make it more robust. The goal of this probabilistic work is to produce data that are as reasonable for statistical uses as possible.

[14] There have also been legal cases concerning different aspects of imputation. For example, in *Utah v. Evans*, the Supreme Court grappled with imputation of count in the creation of the CUF.

The requirement to have a value within scope for every cell shapes the outcome of the CEF. Frustrated by the binary of female/male on the census form, people may purposefully choose both, or neither. The bureau will resolve this to a binary answer for the 2020 Census. Hispanic-origin will also be reduced to a binary, with detailed Hispanic-origin information (e.g., Cuban or Mexican, etc.) being an additional feature. Race is more complicated. Since 2000, people have been able to choose multiple answers for race, in addition to providing detailed race information.[15] During the CEF process, all people are assigned some combination of primary races. They are either one race, a combination of common races, or two or more races. Thus, the data includes information like Asian-only or Black-and-White, but not, for example, American Indian-Black-White; that would be listed as "two or more."[16] Detailed race (e.g., Chinese instead of just Asian) and tribal affiliation have historically been made available as well, but it is not yet clear what level of detailed race will be made available for 2020 data products because publishing this data as it was tabulated in 2010 would have significantly weakened privacy protections.[17]

In the past, after data were cleaned and the CEF was produced, disclosure avoidance was applied via swapping and other noisy interventions to produce the Hundred Percent Data File (HDF).[18] Some households were "swapped" such that whole households were moved from one block to another in order to eliminate the possibility of identifying unique households. Say, for example, that a specific census block had one Black family. In that case, the family is identifiable in the data, as is the other information they provided. Instead of publishing files with such uniquely identifying information, the Census Bureau would sometimes swap that family (and perhaps other families) with a family in another block.[19] The theory with swapping was that one could

[15] Six OMB-level categories dominate: 1) American Indian or Alaska Native (AIAN); 2) Asian; 3) Black or African American; 4) Native Hawaiian or Other Pacific Islander; 5) White; 6) Some other race. In 2010, the decennial form included "some other race" but the ACS form did not. The Census Bureau conducted research to modernize these racial characteristics. They considered adding a Middle East North Africa (MENA) category and collapsing the Hispanic-origin and race data. The Secretary of Commerce did not pursue these changes when he submitted the 2020 census questions to Congress.

[16] Race and Hispanic-origin data first appear in the PL94-171 and are heavily used during redistricting in order to comply with the Voting Rights Act. These data are of utmost importance to map makers, regulators, and civil rights groups seeking to ensure access to representation.

[17] The Census Bureau is examining a technique to model detailed race and tribal affiliation instead of directly estimating it, but the details of this possible solution have not yet been published.

[18] These techniques are discussed in detail in McKenna, Laura. 2018. "Disclosure Avoidance Techniques Use for the 1970 through 2010 Decennial Censuses of Population and Housing," Census Research and Methodology Directorate, U.S. Census Bureau, Washington DC.: https://ideas.repec.org/p/cen/wpaper/18-47.html.

[19] The details of how the swapping algorithm worked are not public. Some information, alongside the ability to undo swapping, is documented in more detail in Mark Hansen's "To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data," *New York Times*, December 5, 2018. In

never be sure if a census record reflects real data for real people in the block, or if it was a swapped record. While the Census Bureau has always engaged in a range of disclosure avoidance processes, it has never published much detail about these efforts, let alone how many cells in different data products this process impacts.[20] Moreover, bureau documentation implies that swapping is not the only method that it uses during disclosure avoidance.

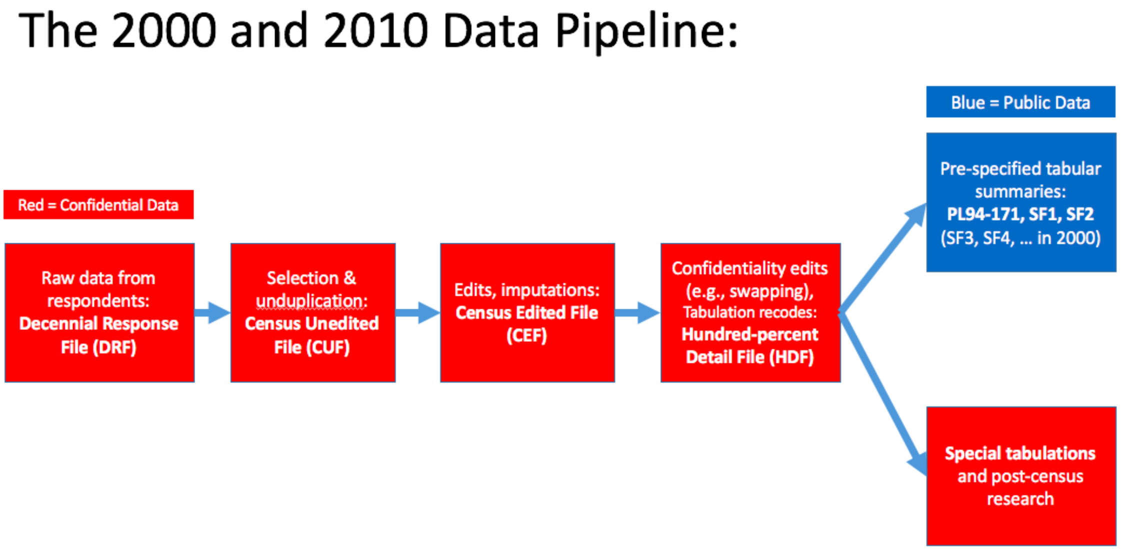
In 2010, the HDF was then sent for tabulation, where the bureau produced the statistical tables that were appropriate for individual data products. At this stage, some tables on certain data products were suppressed because they would disclose too much information.[21] For example, the bureau may have concluded that it could not release tables for geographies with populations under 100. These tables would simply never be published, rendering their exact counts difficult to discern. No counts disappeared from the PL94-171, but the detailed information needed to determine certain subpopulations was not visible in other files.

After the tabulated statistics were produced in 2010, they were then evaluated for quality assurance. The primary purpose of quality assurance was to confirm that the data were coded and calculated accurately and that there were no significant errors in the code. To achieve this, those working on quality assurance produced estimates and evaluated the results against estimates. They were looking for inconsistencies to evaluate whether the inconsistency indicated a problem in the production pipeline or an unexpected change in data.

this story, he describes how the sole residents of Liberty Island (caretakers of the Statue of Liberty) were swapped before describing other aspects of differential privacy: https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html.

[20] For the 2010 census, table suppression rules can be found in the Summary File 2 and AIAN Summary File technical documentation. Other information about disclosure avoidance processes can be found at US Census Bureau. "Disclosure Avoidance and the 2020 Census." https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html.

[21] In particular, Summary File 2 (detailed race and ethnicity) and the AIAN Summary File (tribal affiliations) applied table-level suppression post tabulation.



The 2000 and 2010 Data Pipeline:

Red = Confidential Data

Blue = Public Data

Raw data from respondents: **Decennial Response File (DRF)**

Selection & unduplication: **Census Unedited File (CUF)**

Edits, imputations: **Census Edited File (CEF)**

Confidentiality edits (e.g., swapping), Tabulation recodes: **Hundred-percent Detail File (HDF)**

Pre-specified tabular summaries: **PL94-171, SF1, SF2** (SF3, SF4, … in 2000)

**Special tabulations** and post-census research

*Figure 2 – The pipeline of data collection and processing used for 2000 and 2010 censuses. Source: Modified for clarity from the original graphic developed by the U.S. Census Bureau, 2019.*

After the data were suitably assessed – and the confidentiality of data was protected – the Census Bureau publicly released the privacy-protected tabulated files as specific data products, such as the PL94-171 file for redistricting, Summary File 1 (SF1) and Summary File 2 (SF2) for general use[22], and the American Indian and Alaska Native (AIAN) Summary File for native communities.[23] These were released online and made available to the public via census's website and American Fact Finder.[24] Through this site, data users can download 2010 data product files in different formats. The bureau also produced a range of confidential tabulations for internal purposes that included suppressed tables or other data that could not be released due to the need to protect privacy.

[22] Summary File 1 (SF1) file contains many of the characteristics that are used by demographers and policy makers, including information about all people and households. Summary File 2 (SF2) provides additional detailed race and Hispanic-origin information and offers more nuance for American Indian areas. Details about these files can be found at US Census Bureau. "Summary File 1 Data Set." https://www.census.gov/data/datasets/2010/dec/summary-file-1.html and US Census Bureau. "Summary File 2 Data Set." https://www.census.gov/data/datasets/2010/dec/summary-file-2.html.

[23] For more details on the 2010 data products, see: U.S. Census Bureau, "2010 Census Data Products: United States At a Glance," https://www.census.gov/population/www/cen2010/glance/index.html.

[24] https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

The introduction of differential privacy will change the steps that the Census Bureau takes to produce the statistics that it makes publicly available. The CUF and CEF will still be produced through the same processes as before. But then, rather than applying swapping or previously used disclosure avoidance processes, the CEF data will be sent through the new disclosure avoidance system (DAS), which includes a top-down algorithm (TDA). This will produce a new file called the master data file (MDF), which can then be used for tabulation. Every tabulated file that is produced using this method can then be released publicly.  No tables will be suppressed. That said, because every published table produced contributes to the "privacy-loss budget", the bureau must be careful about what it produces.[25] Additionally, quality assurance now needs to take place both before and after the DAS because after the DAS takes place, it is more difficult to identify anomalies that could indicate data-related errors. At the same time, it is imperative to confirm that the DAS functioned as expected. This alters a range of the bureau's processes.

### When Long-Standing Approaches Stop Working

During the 20th century, the Census Bureau's disclosure avoidance mechanisms grew; more tables were suppressed and more data were swapped.[26] Yet, with the increasing availability of computing power, many of the approaches the bureau took became ineffective at preventing people from triangulating the mountain of census data to identify unique data combinations as part of a reconstruction attack. Furthermore, drastic changes to our contemporary data landscape meant more data about more people was available, which increased the effectiveness of linkage attacks on census data.

A reconstruction attack is a technique that allows computers to triangulate across different statistical tables or data sources to determine which individuals within the data are likely to contain which attributes. The greater the amount and detail of available data, the easier it is to transform a set of statistical tables into a list of individuals with known attributes, such as a person's age, sex, race, ethnicity, and the block on which they live. In other words, reconstruction

[25] All of this will be discussed in detail below.

[26] Cell suppression – that is, the suppression of individual cells in published tables – is used almost exclusively by the Economic Directorate's surveys. The household surveys currently use table suppression, thereby eliminating entire tables when the cost to privacy is too great. Notably, cell suppression also requires complementary suppression so that the details of individual data aren't revealed through the absence in other cells. This has not been a part of the Census Bureau's approach to table suppression.

allows an attacker to rebuild individual records from a set of statistical tables that are purportedly aggregates only.

Reconstruction does not allow the attacker to gain more information than was previously available in the aggregate data. If the statistical tables do not include a person's name, this method cannot provide that information. But if an attacker has access to additional data that includes variables not present in the statistical tables, they can perform a linkage attack on the reconstructed data. In other words, they can match individual records to external data to glean additional information. This is colloquially understood as reidentification, although true reidentification would require being able to confirm the matches, which the Census Bureau can do but an outside attacker could not do unless they approached the individual directly. Still, given the probability that these matches are accurate, a putative reidentification can be understood as being as privacy-breaching as confirmed reidentification.

For some people, reconstruction does not feel particularly invasive. After all, simply knowing that there exists a person who is White, non-Hispanic, 25, and male within a specific block may not mean much. But if this data can be matched with commercial data, it may be possible to know that this person is John Doe. This information could then be used to connect with more information to help construct a digital dossier. While some people see commercial data as more compromising, other people see government data as riskier. Furthermore, census data could contain attributes that people feel are sensitive (e.g., race, household makeup, and citizenship) or could be more easily used to obtain sensitive data (e.g., income and health records, etc.). From the perspective of the Census Bureau, reidentification is a breach of confidentiality – and in violation of the law governing the bureau – regardless of how different stakeholders perceive the sensitivity of specific data.

Reconstruction, linkage, and reidentification attacks have been more theoretical than practical until recently. Starting in the late 1990s and early 2000s, privacy researchers, cryptographers, and security-minded computer scientists started recognizing that advancements in computing would undermine purportedly anonymous open data. Recognizing that statistical tables "leaked" information about individuals, they began showing how it was possible to reconstruct individual records from statistical tables. In 2003, computer scientists Irit Dinur and Kobbi Nissim developed an algorithm that could easily do what was nearly impossible to do by hand: they devised a computationally efficient method for reconstructing individual information from statistical tables.

These developments were disconcerting to members of the Census Bureau. Members of the Research and Methodology (R&M) Directorate began evaluating the vulnerability of census data. In advance of the 2010 census, they believed that they faced limited risk. Still, they understood that every statistical table that is produced leaks some information—and the Census Bureau releases an extraordinary number of tables. The more statistical tables the bureau produces, the more likely that individual records could be accurately reconstructed.

By the middle of the following decade, the technical researchers were concerned that advances in the field might make the 2020 data vulnerable if they continued on their current path.[27] They were also concerned that increased availability of commercial data might make reidentification more feasible. Using the published available statistical tables from only the 2010 decennial census, researchers at the bureau discovered they could reconstruct a complete set of individual records that could effectively serve as a complete public-use microdata file. This is important because the public-use microdata file that the bureau publishes contains only 10% of the population and only with geographies of at least 100,000 people. This attack could provide much more detailed information.

The records that they reconstructed do not fully match the original 2010 CEF because of the earlier confidentiality protections (e.g., swapping), but about 46% of the individual records were perfect matches and, if they relaxed the age to +/- 1 year, about 71% of the individual records were accurate.[28] With those reconstructions, the Census Bureau then attempted to match this data to a subset of commercial data that it obtained; it found that 45% of the reconstructed records matched commercial data. Because bureau researchers are in the unique position of being able to verify their findings with original census data (including names), they were able to confirm that 38% of their matches were correct.[29] In other words, starting with the published 2010 census

[27] Garfinkel, Simson, John Abowd, and Christian Martindale. 2018. “Understanding Database Reconstruction Attacks on Public Data.”*ACMQueue* 16(5): 28-53.

[28] See Abowd, John M. “Tweetorial: Reconstruction-abetted re-identification attacks and other traditional vulnerabilities.” http://blogs.cornell.edu/abowd/special-materials/245-2/.

[29] Abowd, John. M. February 16, 2019, “Staring Down the Database Reconstruction Theorem,” American Association for the Advancement of Science Annual Meeting. https://www2.census.gov/programs-surveys/decennial/2020/resources/presentations-publications/2019-02-16-abowd-db-reconstruction.pdf; and Abowd, John M. and Victoria Velkoff, March 28, 2019. “Managing the Privacy-Loss Budget for the 2020 Census.” Census Scientific Advisory Committee. https://www2.census.gov/cac/sac/meetings/2019-03/managing-privacy-loss-budget-2020-census.pdf. It is also important to note that commercial datasets have a lot more information about many people than the decennial census does. Still, the ability to match decennial data to commercial data raises unique concerns, especially given Title 13.

statistical tables, bureau researchers confirmed that they could reidentify at least 17% of the public by name, which they determined to be a significant loss of confidentiality. Although some of this could be replicated outside of the bureau, few researchers have been willing to attempt doing so out of ethical concerns.

It is quite possible that by purchasing or acquiring even more commercial data, the ability to match more of the data would have been greater – and that will certainly be the case in 2020 – but as a research exercise, this level of matching confirmed the Census Bureau's concerns. While much of the data collected by the bureau are duplicated in commercial data, which is how the matches were possible, the bureau also has a significant amount of data that is not as readily available. The quality of address data in commercial data is wildly variable. Census data has much more significant detailed race and tribal affiliation data. And almost no commercial data has household configuration. Household vacancy is also not readily available commercially in many geographies. Moreover, because commercial data has other sensitive attributes, combining census data with commercial data can be doubly revealing, strengthening the quality of commercial data, which benefits commercial entities. Additionally, since the Census Bureau integrates a range of administrative data into its systems, the data that are exposed are the combination of what people self-report and what is found within other government records. It is precisely the sensitivity of these combined data that make academic researchers loathe to replicate the Census Bureau's self-attack, but the proof-of-concept attack combined with the lack of legal protections and the wide array of unethical domestic and foreign actors, is disconcerting to those who are familiar with these attacks.

Because the reconstruction and linkage attacks prompted anxiety among both those in academia and those at the Census Bureau, computer scientists began to focus on how to respond to these vulnerabilities with new protections. In 2006, computer scientists Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith first proposed a new way to mathematically assess and guarantee the level of individual information that could be reconstructed from statistical tables. This method became known as differential privacy and was a mathematical definition for "guaranteeing privacy."[30] This concept opened up an entire field as researchers began looking for different ways to implement efficient systems for producing statistical tables that could be evaluated through the mathematical guarantees of differential privacy. In other words, because

[30] In discussions of differential privacy, computer scientists speak of privacy in a way that is similar to how census advocates speak of "confidentiality." What's at stake in terms of privacy in this context is whether someone can discern features about individuals from aggregated data. A "guarantee" means that there is a mathematical proof that stands behind the measurement that is conveyed.

statistical tables leak information and because leakage is a vulnerability, these researchers developed techniques to evaluate how much leakage existed—and a process for evaluating how "noise", that is, uncertainty in the form of mathematical randomness, could be introduced to reduce the leakage.

The Census Bureau began exploring differential privacy as a potential avenue for addressing reconstruction attacks on its data as early as 2006 and started publicly implementing differentially private tables for select statistical products starting in 2008.[31] As a statistical agency with researchers who are advancing knowledge on many fronts, the bureau consistently experiments with new statistical and computational techniques to advance its work. When it began experimenting with differential privacy, it did not believe that the vulnerabilities would advance as quickly as they did. For the researchers inside the bureau, their ability to perform a significant reconstruction and reidentification attack was a huge wake-up call. Recognizing how much commercial data had increased in the 2010s and, thus, how much more would be available by 2021, they were confident that reidentification would only get easier. While external researchers might not be able to affirmatively verify the matches, bureau researchers knew that many of the reconstruction and reidentification matches were overwhelmingly accurate and would only become more so.[32] In effect, it became clear to bureau researchers that protecting the 2020 census with the methods used in 2010 would yield a much higher reidentification rate than the 17% found with the external data available in 2010. Thus, they began to invest heavily in developing systems that could abide by differential privacy.

Differential privacy works to prevent accurate reconstruction attacks while making certain that the data are still useful for statistical analyses. It does this by injecting a controlled amount of uncertainty in the form of mathematical randomness, also called noise, into the calculations that are used to produce data products. The range of noise can be shared publicly because an attacker cannot know exactly how much noise was introduced into any particular table. With differential privacy it is still possible to reconstruct a database, but the database that is reconstructed will include privacy-ensuring noise. In other words, the individual records become synthetic byproducts of the statistical system.

[31] The Census Bureau was the first organization to release data protected by differential privacy. In 2008, they released the On The Map tool. See: https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html.

[32] When a researcher knows that their match has a 99% likelihood, it is effectively "beyond a reasonable doubt," even if the Census Bureau does not formally confirm the match.

There are different ways to build a system to protect the confidentiality of census data under differential privacy. The Census Bureau opted to focus on a top-down approach to maximize utility for redistricting and funding allocation.  By understanding how data users might cross-tabulate variables or otherwise conduct statistical analyses, noise could be added to cells in the tables such that it would be difficult to discern individuals. To do this well requires understanding how different statistical tables interact with each other. A top-down approach can produce more tailored statistical tables because noise is injected in ways that preserve the most important statistical efforts. If race is the most important variable, more noise can be introduced to the age variable to preserve the accuracy of the race variable. If one table is more important than another, more noise can be provided to the less-frequently-used table than the one that's more essential. While this control is an asset, it also introduces significant governance and policy implications.

Noise is introduced to statistical tables to reduce the leaked information. Within the construction of a differentially private system, it is necessary to make choices about how much noise is tolerable and how much risk, or total privacy loss, is acceptable. The creators of a differentially private system set a maximum cumulative privacy-loss budget. This is defined by the variable epsilon.[33] Epsilon can be conceptually understood as a set of knobs. Dial the knobs one way to create higher levels of noise but lower levels of confidentiality risk. Dial them the other way and the data are more accurate, but the risk to confidentiality (and, therefore, "privacy loss") is higher. The system has a global privacy-loss budget, but each geographic level and table also has a local privacy-loss budget that must be managed such that the interaction of all tables does not result in leakage that exceeds the global amount. The actual local budget is a combination of the particular table and geographic level. Because of these locally defined epsilons, it is possible to introduce more noise to a specific variable (e.g., sex) than others (e.g., race).

Differential privacy implies a set of interconnected trade-offs. The lower the amount of noise injected into one particular table, the greater the accuracy of that table. Greater accuracy of one table means less accuracy is available for other tables because the total epsilon, or privacy-loss budget, is fixed. Once the privacy-loss budget is determined, the available accuracy must be shared among all the published tables. Increasing the accuracy of some tables without reducing the accuracy of others can only be accomplished by increasing the total privacy-loss budget and, therefore, increasing the risk of confidentiality violations. New tables do not inherently increase the privacy loss. For example, the first large file that the Census Bureau produces – the PL 94-171

[33] *Epsilon* sounds fancy, but it's just a Greek letter: $\varepsilon$. Mathematicians and computer scientists regularly use Greek letters as variables in their notation.

– is going to leak a lot of information. But if a new data product has significant overlap with this data and just introduces one new variable – say, sex – the amount of leakage introduced by the second data product is not nearly as great.

Setting epsilon as a mathematical variable requires making a numerical choice between ~0 and infinity.[34] Technically, this decision could be revised over time to reduce the noise if that becomes important.[35] De facto, all census data will be made available 72 years after the census when the data are all made public by law. Once the global epsilon is set, decisions need to be made about how to "spend" that budget based on the privacy loss of every statistical table that will be produced. These decisions have significant consequences because they govern the trade-off between confidentiality and accuracy, which affects data utility. The data must be fit for purpose, and determining those purposes must be done upfront.

### A Fundamental Rupture in Data Product Production

After recognizing the threat to confidentiality from reconstruction and linkage attacks, the Census Bureau evaluated its options. The default would be to drastically reduce the number of data products produced and significantly increase the amount of swapping and table suppression. With differential privacy, it saw an opening to shift paradigms and still make data available so it set about to build a disclosure avoidance system (DAS) based on differential privacy.

To build a DAS that is differentially private, the team building the system must identify all statistical tables and unique data cells that will be published in order to assess the total privacy-loss cost of each table before production. Unlike the other techniques that the Census Bureau implemented in the past, a formally private approach demands critical decisions about the cumulative privacy-loss budget must be decided before statistical tables are produced. Decisions concerning one table affect the production of future tables because the data are interdependent and the combination of tables is what produces the risk to confidentiality. Moreover, although a top-down approach provides greater control over how the noise is distributed, it also requires decisions be made about how to distribute the noise across the system.

[34] Epsilon cannot actually be zero, but it can be the smallest positive non-zero number that one could imagine.

[35] While epsilon can be increased over time, it's not possible to decrease epsilon and increase privacy because the information has already been leaked.

As I described above, the 2010 data products were the end of a clear pipeline:

> CUF→CEF→swapping and disclosure avoidance→HDF→tabulation→table suppression→data products

The 2020 differentially private approach changes this pipeline in subtle but important ways:

> CUF→CEF→noise-injection DAS→MDF→tabulation→data products

Because the noise that is introduced by the DAS protects the confidentiality of the data under the guarantees of differential privacy, the master data file (MDF) is not sensitive. Unlike the HDF, the MDF does not necessarily need to be kept confidential because it is the file that would be reproduced if an attacker reconstructed the data from published tables.[36] That said, the production of the MDF by the DAS depends on the bureau knowing what kinds of data products need to be produced. The other major advantage of this approach is that special tabulations do not need to be preserved as sensitive because the data behind them are not sensitive. At the same time, the ability to produce any and all tabulations is limited by the privacy-loss budget, or epsilon.

[36] The MDF limits the ability to match against commercial data because the resultant data subjects are not actually composed of real people. That said, even an artificially constructed composite of a person could be matched against commercial data. After all, it's reasonable to imagine that there's one person living in a Manhattan block that's White, non-Hispanic, male, and 40. To understand these possibilities, the Census Bureau intends to replicate its linkage attack and reidentification analysis to understand what percentage of people could be identified this way.

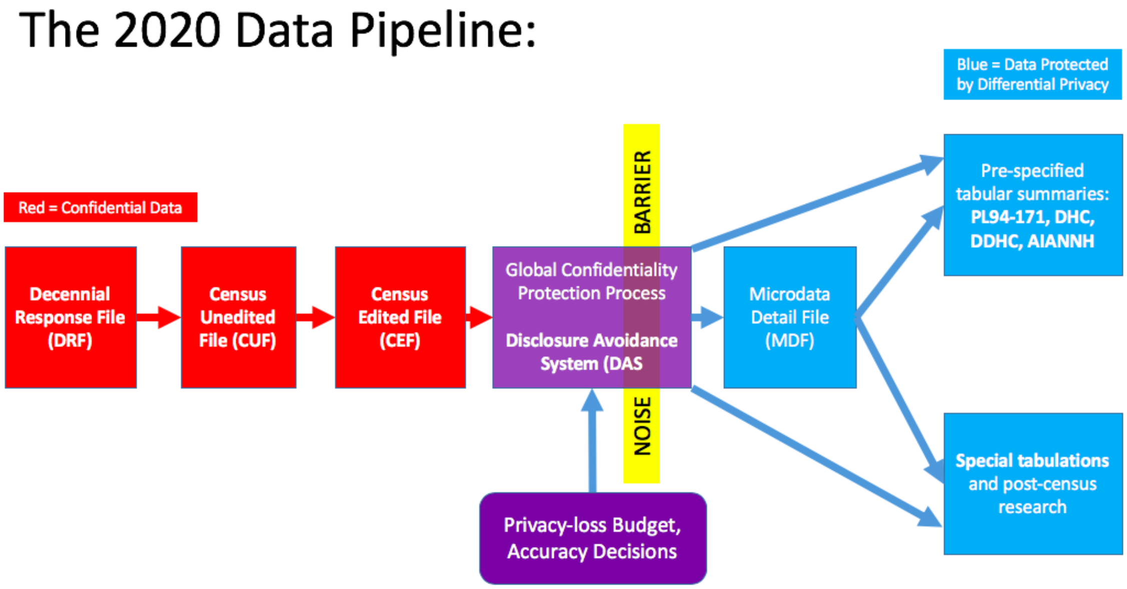

The 2020 Data Pipeline:
Blue = Data Protected by Differential Privacy
Red = Confidential Data
Decennial Response File (DRF)
Census Unedited File (CUF)
Census Edited File (CEF)
Global Confidentiality Protection Process
Disclosure Avoidance System (DAS
BARRIER
NOISE
Microdata Detail File (MDF)
Pre-specified tabular summaries: PL94-171, DHC, DDHC, AIANNH
Special tabulations and post-census research
Privacy-loss Budget, Accuracy Decisions

*Figure 3 –The proposed pipeline for producing 2020 census data products. Source: Modified for clarity from the original graphic developed by the U.S. Census Bureau, 2019.*

To implement differential privacy, two major questions must be determined before the DAS is run: the global privacy-loss budget (or global epsilon) and the allocation of the budget across all of the tables and features (local privacy-loss budget allocation). This means determining in advance all of the data products that will be produced. If the bureau were only concerned with producing the one constitutionally required piece of data – the resident population of the 50 states – none of this would be necessary because that information comes from the CUF. However, additional statutes and regulations place demands on the privacy-loss budget for the 2020 Census. The data requested by the states for redistricting purposes are especially detailed, with information reported at the block level. Further, there are a range of other data products produced that are widely used by federal, state, and local agencies, as well as researchers and the business sector; each of these would need to be considered.

In choosing how to implement the system, the DAS team also needs to consider what kinds of analyses are critically important so that aggregated data are less noisy where they matter.  The bureau's top-down algorithm is designed such that noise is introduced from the top of the geographic structure down. Noise is first introduced at the state level (for features) and the county level (for counts). Each subsequent geographic level down the spine receives additional noise, such that the cumulative noise at the block level is the greatest within the system. The goal is that when data users add up blocks along the census geographic spine, the data will become less noisy.

In deciding how to build the system and protect privacy, the DAS team must also factor in invariants. Invariants are data that cannot be altered; thus, there is no noise and these reflect the original count.[37] Under all disclosure avoidance approaches to date, there have been invariants. While there has been no official confirmation of all 2020 invariants, the Census Bureau has confirmed that the state-level total population counts will be invariant, which guarantees that if a state is told that its population is 4,339,367, then that's what it is according to the best information that the bureau has. They have also decided to make certain that all of the counts down the spine will add up to the state-level number.

The DAS team has also implemented a set of constraints on implementation. In an effort to not confuse different stakeholders (e.g., policy makers, lawyers, and some data users) or introduce legal complications, the bureau decided that all cells must contain positive whole numbers. This means that there will not be any blocks with -5 or 4.27 people in them. Negative population counts and fractional people can be useful for statistical purposes, but would be unlikely to make sense to many members of the public.[38] Yet, these choices have consequences. To add noise to small blocks without going negative requires that small blocks, on average, get bigger. In turn, because of state-level count invariants, this means that the largest blocks, on average, get smaller. On the plus side, information about these ramifications can be published so that statisticians can adjust their models accordingly.

The more constraints or invariants that the bureau introduces, the more difficult it is to produce differentially private data that is not especially noisy in another area. Consider, for example, the challenge of producing statistics about housing unit vacancy rates. If the bureau makes key features of the MAF invariant, this means that the number of housing units and the presence of GQ on each block must be exact. Thus, all of the noise must be introduced to other two key pieces of additional information about housing units: whether they are vacant and, if occupied, whether they are owned or rented. If high-quality data about vacancy rates are important, something else

[37] There are no statutorily required invariants, although the bureau has long maintained certain categories as invariants (e.g., voting age population at all geographic levels and number of housing units at all geographic levels). See: Abowd and Velkoff. 2019. "Managing the Privacy-Loss Budget."

[38] Additionally, some local statutes require districts whose size does not differ by more than 1 person. Having fractional people would trigger legal debates about whether 1.05 people is acceptable and if so, whether a 1.51 difference would be. Negative people would also produce confusion in these situations.

must give. One approach would be to remove the restriction about the number of housing units; another approach would be to restrict the lowest level of geography.

Introducing differential privacy into the DAS creates a number of different ruptures when producing and disseminating census data. This approach changes many internal Census Bureau processes. Because the DAS must happen before tabulation, the mechanisms for quality control of the census data inside the bureau must change. Moreover, there is less experience and familiarity with debugging this kind of system – or the ability for subject-matter experts to easily spot a problem with the tabulation process – when noise is intentionally introduced. The bureau must also change other parts of its processes that depend on census data, including how it produces population estimates and how it manages the production of social science research that takes place within the bureau. This is necessary so that new products and new research do not leak additional information that could compromise the confidentiality of the data.

More importantly, a turn to differential privacy also requires determining upfront what types of data analysis are needed and what should be prioritized in deciding how noise is allocated. This requires that data users and the Census Bureau work together to make decisions.

### Speaking Past One Another

In December 2018, the Census Bureau formally announced its intention to implement differential privacy as part of its 2020 disclosure avoidance processes. This announcement was met with confusion and frustration on the part of many data users and census advocates.[39] Many census advocates had never paid much attention to earlier disclosure avoidance processes and did not feel equipped to understand what the changes meant. Meanwhile, many data users felt blindsided, frustrated that they were not consulted over a change this significant. Some dismissed the risks that the Census Bureau articulated in justifying the transition.[40] Others wanted more

[39] The arguments made in this section are based on conversations with both census data users and people at the Census Bureau. I spoke with dozens of people in the fall of 2019 in an effort to understand the disconnects between different stakeholders.

[40] Those who challenge the introduction of differential privacy often argue that the Census Bureau had the option to continue with its current approach. Usually, these data users argue that the risk to confidentiality is not as great as the bureau suggests it is. Others suggest that Title 13 is outdated and should be changed. Still others interpret the law in a different way than the bureau does. Regardless of these external arguments, bureau leadership believes that they would be violating the law and failing to uphold their oath to confidentiality if they were to proceed in a business-as-usual fashion.

information than the bureau was prepared to offer. Frustration grew as bureau professionals and data users failed to effectively communicate with one another.

Once the bureau announced that it would integrate differentially private methods into its disclosure avoidance procedures, it published a Federal Register Notice (FRN) to solicit feedback about what data products and fields were most critical to data users in light of the transition. This triggered a range of negative reactions from a wide variety of data users who were concerned and confused. What they took from the FRN was that the Census Bureau did not intend to publish many of the tables it had published in the past and was not communicating which data it would be publishing. From the perspective of those working on the DAS at the bureau, this was an attempt to engage the data user community in making critical decisions. They understood that a differentially private approach would require them to determine all data uses in advance, but the data user community had never operated this way. The 2018 FRN was the first time that many data users were introduced to the concept of differential privacy. Rather than feeling included in the process, many felt as though they were being excluded from decision-making that would significantly affect their work.

The bureau did not educate data users in advance of the FRN or guide them about how to effectively respond. Some data users tried to respond to the request as best they could, articulating what they did with what data. Yet because most data users rely heavily on the American Community Survey (ACS), many emphasized ACS data even though that was out of scope for the FRN. Others simply requested that they be able to get access to all of the tables and variables that were available in 2010. This was not what the bureau expected would happen.

By this point in the process, the disconnect became increasingly visible. Presentations made at various census-related advisory events and academic conferences often escalated confusion or discomfort. Attempts by the Census Bureau to engage data users on key components to implementing differential privacy were met with resistance because the data users did not share the view that a move to differential privacy was necessary. The lack of an education campaign and the perceived speed of this transition left many data users feeling as though they had not been consulted on – and did not have enough time to digest – such a significant change; they believed that the bureau was making a unilateral decision without fully understanding the impact this would have on the broader data user community, from the map makers to the academics, the city demographers to the policy makers.[41]

The bureau faced resistance to its strategy, which it was not prepared to address effectively. Census Bureau leadership had hoped for cooperation in the decision-making process. Rather than working together to solve this complex problem in a politically fraught context, data users and the bureau's DAS team failed to find common ground early on. One challenge stemmed from the disconnect between how census data users and the computer scientists building the DAS conceptualize the production of data products. Another stemmed from the cultural logics and communication styles of these two groups.

Data users are accustomed to having a significant level of access to census data, with an ever-increasing amount of data available to them each decade. In recent years, the Census Bureau has made census data readily available through an online portal. Even before that, a range of services have been made available to data users. One of the most commonly used is IPUMS,[42] which the Institute for Social Research and Data Innovation at the University of Minnesota hosts. Another is the Inter-university Consortium for Social and Political Research (ICPSR), which is hosted in the Institute for Social Research at the University of Michigan. When data users work with census data, they are not required to tell the bureau what they are doing, nor is the bureau structured to learn about all of its users 'activities. The statistical agency only has clear visibility into the proposals that users put forth to work in the Federal Statistical Research Data Centers, where data users are required to pass background checks and be approved to work on certain projects. Yet this is a small fraction of census data users, in part because most data users do not need this level of access. The bureau asked census users and stakeholders to explain what they did with data, they were asking because the existing data publishing model limited their insights into data user practice.

The Census Bureau is comprised of researchers from many different disciplines. The team dedicated to differentially private disclosure avoidance is predominantly computer scientists and economists. Their approach to building the DAS is rooted in optimizing specific statistical outcomes. They value transparency, which in their world means making the code available for

[41] One of the first such public critiques is: Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. "Differential Privacy and Census Data: Implications for Social and Economic Research." *AEA Papers and Proceedings,* 109: 403-08. https://www.aeaweb.org/articles?id=10.1257/pandp.20191107.

[42] IPUMS once was an acronym, but is now the proper name. See: https://ipums.org/what-is-ipums.

inspection. For most data users – including demographers, statisticians, and social scientists – transparency means having access to reliable data which, when combined with well-established statistical methods, can help them understand patterns in census data. They are not looking to analyze the DAS code. Thus, when the DAS team released its code alongside the now-public 1940 data to show its work and invite feedback, it believed it were doing outreach. What data users experienced was poorly documented code, incomplete guidance on how to apply the code, and inconsistent support from the bureau.

Most data users were ill equipped to work with the DAS code, but some did their best with what they had been given. Once they started working with it, however, they asked different questions of the system than the DAS team had expected. Rather than exploring the code, those who decided to invest in this process wanted to know how different the data would be and how such data might affect their scientific work. They produced histograms comparing the data at different levels of epsilon. What they saw concerned them. When bureau scientists saw the data users 'analyses, they grew frustrated because they did not see them as a sensible way to evaluate their system. Again, data users and the DAS team spoke past one another. Even when conversations took place, data users and bureau scientists struggled to discuss the types of trade-offs that the bureau felt it needed. Meanwhile, the very idea that the bureau was moving forward without these tensions being resolved gave data users the impression that the bureau was not listening.

Throughout this process, a communication chasm grew in every direction. Census Bureau representatives thought that they were being transparent and open in their communications, but data users were at best confused by what was being said and more often felt as though the communications were condescending or arrogant in tone. Many data users felt as though they were sharing their concerns but grew frustrated by what they experienced as sporadic and ad-hoc outreach. Bureau representatives felt overwhelmed, understaffed, and frustrated by what they were hearing. Those on the DAS team did not know how to interpret data users ' analyses and others 'feedback in light of the trade-offs they felt they needed to make. There were also disagreements within the bureau itself, and disagreements among data users spilled out into academic conferences, at census-related events, and in news stories. Meanwhile, a group of census stakeholders focused on the GOTC campaign grew increasingly worried that these conversations would explode, add to the fear and doubt in the Census Bureau, and hinder GOTC efforts, especially if the bureau's efforts were misinterpreted in news media coverage in ways that suggested a potential loss in confidentiality. Guaranteeing confidentiality is a critical component of GOTC efforts, and stakeholders 'preliminary outreach efforts suggested that many people already doubted confidentiality due to distrust in the government.

The Census Bureau started taking steps to assuage concerns. First, it announced that differentially private methods would not be applied to the ACS before 2025 at the earliest and that data users would be consulted in the process.[43] It did so because many data users were concerned that the unilateral rollout of the new DAS for the decennial was indicative of what would happen to the ACS, which concerned many more data users.

Second, in collaboration with the Committee on National Statistics (CNSTAT), bureau officials agreed to republish some of the 2010 data products with the new DAS so that the data could be more meaningfully compared. This was important because, even if the publicly released code was similar, the application of the 2020 code to the 1940 data created a false comparison for the issues that were of interest to data users. The 1940 data were not representative of contemporary census data: there were no block-level geographies; the categories used for race were different; the population was much smaller; and the distribution of the population over the country was significantly different.

Using the 2010 CEF and the 2020 DAS code, the bureau published a new PL 94-171 file and a portion of the SF1 file, excluding household joins and detailed race characteristics. [44] This is still an imperfect comparison because the public 2010 data had already been altered by prior privacy protection mechanisms, like swapping, table suppression, and other disclosure avoidance techniques. As of this writing, some data users are diligently working to understand and compare these results. The bureau is expecting to hear the first wave of findings from these experiments at a meeting of CNSTAT on December 11-12, 2019.

While the feedback from the CNSTAT evaluations will inform the Census Bureau's processes, those who are involved in this effort are primarily academic and statistical data users. City demographers and civil rights groups are trying to rally additional data users to participate in this evaluation, but there are limited resources available to these groups to support these efforts.

[43] What consultation looks like has remained vague and will certainly be a key factor moving forward.

[44] Most tabulated files produced by the census focus on individuals (e.g., the number of people of a certain age, race, or sex). Some tabulated files center on households (e.g., the number of households with two children). To produce these tabulated files, census data must be "rejoined" with housing data. These are discussed as "household joins," or cross-tabulation of household relationship characteristics.

Many of the data users that will depend on the census data have not even started considering, let alone accounting for, how disclosure avoidance might shape their practice. Some data users are simply waiting to see what happens before engaging. Others lack the time, funding, or capacity to invest in doing assessments with incomplete information. Still others feel as though the published numbers are what is important, regardless of what noise has been infused. Many in the redistricting community are more concerned with how the differentially private data might affect their legal claims; they must work with the data that the bureau publishes to produce maps that comply with a range of legal requirements, most notably that the districts must be equitable and must comply with the Voting Rights Act. It is unclear how courts might respond to claims that the noise justifies certain decisions in drawing districts. Those who use the data to comply with federal funding laws are also not sure how their processes will be affected by changes in the release of data, let alone how the data will be treated when disputes need to be resolved.

While the move to differential privacy is critical to protect confidentiality, how it has unfolded – and continues to unfold – introduces risk to the execution of the census. In short, as we sit on the cusp of 2020 and the GOTC efforts start to unfold, the Census Bureau needs to be able to guarantee confidentiality *and* data users ’need to feel as though they will get what they require to do their work. This is a tall order in the current climate.

### Opening a Pandora’s Box

The move to differential privacy has also destabilized how the quality of census data are evaluated and perceived. For some census data users, the introduction of differential privacy is not seen as a mechanism of disclosure avoidance but as an active sabotaging of the census data that undermines improvements in data quality each decade. In short, some see the choice to purposefully introduce noise as a challenge to the validity of the data.

Disclosure avoidance has not been deeply interrogated in previous censuses. Discussions of differential privacy have increased the visibility of disclosure avoidance procedures.  Far from being reassuring, a closer examination of disclosure avoidance and statistical imputation methods has triggered anxiety in data users about the underlying quality of the data more generally. Fundamentally, census data are legitimate because we collectively believe them to be good enough for use in our democracy. The Census Bureau defines quality data as that which are “fit for use,” meaning that they are good enough to serve the statutory purposes. Watchdogs and members of

Congress hold the Census Bureau accountable for its processes and push the agency to take active measures to count everyone to improve quality. Yet the data never were and never will be perfect.

In each and every census, some people have not been counted and some people have been counted twice or included erroneously. Omissions have played a significant role in discussions about the quality of census data, as have differential undercounts between different communities.[45] Each census faces different challenges.

After each census, the Census Bureau attempts to assess the quality of its work.[46] For example, it conducts a post-enumeration survey and uses alternate data sources to produce an independent measure of the census year population called Demographic Analysis. The bureau repeatedly develops new techniques for identifying limitations to its data.[47] All of this work is done to improve future censuses and promote transparency.

At the same time, there are certain persistent problems. Whole households are missed due to operations failures, and individuals within households are missed for a range of reasons, including families not wanting to report some members of the household or not knowing that they should. There are also problems with getting the features of individuals right. Some census respondents intentionally provide deceptive or incomplete information for any number of reasons. Others identify in ways that do not align with the categories provided by the census, and question the value of a census that does not recognize them on their terms and categorizes them into officially accepted categories. There have also been enumerators who have not been truthful and people who have not had enough information to accurately respond to what the census asks. These are just some of the data collection challenges.

Many types of error that are in the census data have declined over decades. The census used to be tabulated by hand; computers are more reliable, though current paper forms are not always

[45] In 1947, Daniel Price used selective service data to argue that there was a significant “under-enumeration” in the 1940 census. His work showed a differential undercount; Whites were undercounted, but at a rate far lower than that for Blacks – especially Black men. Decades later, this prompted civil rights advocates to push for eliminating the differential undercount, which has become the dominant frame for understanding the quality of each subsequent census. Price, Daniel. 1947. “A Check on Underenumeration in the 1940 Census.” *American Sociological Review*, 12(1), 44-49.

[46] For a thorough overview of quality assurance processes, see: Hogan, Howard, Patrick J. Cantwell, Jason Devine, Vincent T. Mule Jr., and Victoria Velkoff. 2012. “Quality and the 2010 Census.” *Population Research and Policy Review* 32(5), pp. 637-662.

[47] In tabulating same-sex couples after the 2010, census researchers recognized that some percentage of respondents appear to mark their sex inaccurately, thereby creating a situation in which some opposite-sex couples are marked as same-sex couples. See http://socialcapitalreview.org/wp-content/uploads/2012/05/sshs2010c.pdf.

processed accurately by machines. Census enumerators are no longer patronage jobs, which has reduced earlier types of bias. Missing or incomplete records have always had to be addressed, and the processes for doing imputation and assignment have improved with the increasing availability of administrative records.[48] Still, not everyone is present in administrative records—and administrative records are subject to error.

After each census, data users also attempt to understand the quality of the data in order to push the bureau to do better. Data users are concerned that the same techniques that are designed to prevent breaches of confidentiality will also undermine their ability to assess limitations in the data. For example, New York City's demographers identified low-performing supervisors through analyses of anomalous vacancy rates.[49] Advocates for children actively challenged the bureau to pay attention to the undercount of young children (under age 5) and produced extensive research on the problem, highlighting which children were most likely to be missed (e.g., those living in complex households) and also describing which children were more likely to be counted twice (e.g., those of divorced parents). This advocacy has informed the bureau's 2020 operational plan,[50] unlocked philanthropic support, and galvanized census stakeholders.

Because the Census Bureau has consistently worked to develop new methods for addressing limitations to its data and operational procedures, data users have typically accepted that the data are the best that they can be. Those who have been deeply invested in understanding the details of the census know that disclosure avoidance mechanisms have always introduced some noise into the system, but the bureau has always maintained that the noise purposefully introduced is very small compared to the noise that comes from mistakes, imputation, and other forms of human error. Data users have accepted – or, in some cases, not known about – this reality. The data that have been produced for the census have not included margins of error, although imputation rates

[48] The Census Bureau uses a wide range of administrative records collected and maintained by other governmental agencies (e.g., birth records or tax records) in its operations. While the bureau does not share its individual-level data with other agencies, other agencies share data with the Census Bureau in order to help with its mission. The bureau uses administrative records during the data collection phase to better understand who it may not have counted. After data collection is complete, it also uses administrative records to improve the quality of the data. These data helps with correcting errors, filling out missing information ("assignment" and "imputing"), identifying duplicates, and assessing quality, among other things.

[49] Salvo, Joseph J. and Arun Peter Lobo. 2013. "Misclassifying New York's Hidden Units as Vacant in 2010." *Population Research and Policy Review* 32(5): 729-751.

[50] US Census Bureau. February 2019. "Investigating the 2010 Undercount of Young Children – Summary of Recent Research." https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/final-analysis/2020-report-2010-undercount-children-summary-recent-research.html.

have been made available and show variance across different demographics.[51] With this much attention on the new disclosure avoidance processes, one open question is whether the data users will accept the data as good enough.

### Moving Forward, Managing Unknowns

Throughout the 20th century, the Census Bureau has strived to balance the needs and interests of data users alongside the confidentiality of the information collected, recognizing the importance of public trust in the data collection process. Fundamentally, the loss of trust (which could be easily spurred by a breach of confidentiality) would undermine robust data collection, rendering the data products less accurate and, in some cases, useless. At the same time, the bureau has a public-facing mandate. At the very least, it is committed to providing data to the users defined by statute who contribute to redistricting, the expenditure of federal funds, and the implementation of federal policies, such as civil rights laws. Yet as a statistical agency deeply committed to the scholarly pursuit of advancing knowledge, it has a tacit and explicit commitment to a much broader range of social scientists as well.

The bureau is a large bureaucracy populated almost entirely by career civil servants—a structure that helps build resiliency to political pressure that could undermine scientifically driven decision-making. But it is also a government agency that is expected to operate transparently and be accountable to the Office of Management and Budget, the Commerce Department, and Congress. The Census Bureau complies with administrative law through FRNs, holds public meetings, and convenes advisory groups to guide its processes. It produces detailed operational plans, outlining its decisions and plans to conduct the census. The R&M Directorate, which is building the DAS, exists to help ensure that the techniques and operations that surround the census are as sound as possible. It is within this context that the risks to confidentiality were identified and the remedy of introducing differential privacy was proposed.

The effort to conduct the decennial census and the parallel civil society, nonprofit, and community-based GOTC efforts are fully underway. Within the bureau, there are tightly orchestrated processes, strict timelines, and procedures for evaluating the various activities that are underway. Many of the decisions about what to do and how to do it were based on

[51] U.S. Census Bureau. February 2019. “2010 Decennial Census: Item Nonresponse and Imputation Assessment Report.” https://www.census.gov/library/publications/2012/dec/2010_cpex_173.html.

experiments conducted earlier in the decade and were finalized and tested long ago, although some critical operational tests were curtailed due to congressional funding restrictions. Part of what makes census stakeholders and data users anxious about the implementation of differential privacy is that it is not finalized. There are many unknowns. The timeline is not set. The production code is not finished. Epsilon is not nailed down. Only some data products are publicly known. In many aspects, similar details were not hammered out by this time in 2009 (and there was never a production-level DAS built), but that is not reassuring for many because differential privacy feels like more of a significant change.

At present, there is one overarching question: *What is the best way to maximize data utility while protecting confidentiality?*

From my vantage point, weakening the promise of confidentiality undermines the ability to count the hardest-to-reach populations. For this reason, protecting confidentiality is critical. Right now, the best technique to reduce the risk of exposure is rooted in differential privacy, even though this will require a loss of some data. But something has to give. There is no way to have full confidentiality *and* complete accuracy. The question is how to relax the constraints.

One possibility would be to reduce the geographic precision. For most of its history, the Census Bureau did not produce block-level data and these data were not needed or available for redistricting in significant parts of the country. While census blocks were first introduced in 1940 for major cities, full block-level information in all sparse geographies was only available starting in 1990.[52] While block-level detail is desirable for many people and for many reasons, the cost to confidentiality in the modern age is growing and, in my opinion, is too great to risk. Should the Census Bureau not publish block-level data, there would be much greater flexibility in the privacy-loss budget, allowing for much higher levels of accuracy elsewhere in the data.

Unfortunately, the Census Bureau has been given orders by the Secretary of Commerce and many states to produce block-level data for many variables. It has been given orders to produce block-level housing unit counts, race and Hispanic-origin data, citizenship, and group quarters counts. The Census Bureau is under orders to produce both the PL94-171 and CVAP redistricting files are ordered to be produced at block-level. Stopping the block-level production of these data would require an act by Congress, which makes this a radical recommendation. It would also require

[52] Small-area data is relatively modern. See: US Department of Commerce. 1994. "Census Blocks and Block Groups." *Geographic Areas Reference Manual*. https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf.

a wide universe of data users and Congress to find common ground and collectively agree that these data are not essential.

At present, the bureau is stuck between a rock and a hard place. It is tasked with balancing confidentiality and data utility but has very limited levers to work with. Completing the task requires trade-offs, which will be unlikely to please anyone. Status quo is not an acceptable solution; there is no returning to a world in which data products are published without protections, and the timeline is shrinking to complete the work needed to produce the data. At the same time, the disconnects between the Census Bureau and its data users cannot be ignored; the bureau needs to better understand how its stakeholders use its data, and fast. As the decade ends, it is imperative that all who are invested in the census and its data must actively cooperate and compromise to find a tolerable outcome that is good enough. Everyone involved is working with constraints, such as not enough resources, not enough information, and a need for compliance. Still, this puzzle must be unlocked and the community as a whole must devise a plan that works within existing constraints. What's at stake isn't simply the availability of the data; it's the legitimacy of the census.

## Author Biography

danah boyd is the founder and president of Data & Society, a partner researcher at Microsoft Research, and a visiting professor at New York University. Her research is focused on making certain that society has a nuanced understanding of the relationship between technology and society, especially in light of how inequalities are reinforced through sociotechnical systems. She is the author of “It’s Complicated: The Social Lives of Networked Teens” and has authored or co-authored numerous books, articles, and essays. She is a director of the Social Science Research Council, a trustee of the National Museum of the American Indian, and a director of Crisis Text Line. Originally trained in computer science before retraining under anthropologists, danah has a Ph.D. from the University of California at Berkeley’s School of Information.

## Acknowledgments

This paper was made possible by spending countless hours with differential privacy experts, census data users, and census advocates who have thoughtfully challenged me to better understand and communicate the Census Bureau's goals and plans and the frustrations of data users. While not everyone appreciated public acknowledgment, I want to thank some of those who helped me improve this document along the way: John Abowd, Margo Anderson, Dan Bouk, Norin Dollard, Cynthia Dwork, Maria Filippelli, Abraham Flaxman, Kyla Fullenwider, Simson Garfinkel, Matt Goerzen, Daniel Goroff, Mark Hansen, Ron Jarmin, Charley Johnson, Jae June Lee, Terri Ann Lowenthal, William O'Hare, Ken Prewitt, Deborah Stein, Whitney Tucker, David Van Riper, and Lars Vilhuber. Their contributions to this living document do not imply that they agree with my analyses or depictions. I also want to thank Patrick Davison, Rati Bishnoi, Yichi Liu, and Sareeta Amrute for their contributions to producing this living document.

## DATA & SOCIETY

Data & Society is an independent nonprofit research institute that advances new frames for understanding the implications of data-centric and automated technology. We conduct research and build the field of actors to ensure that knowledge guides debate, decision-making, and technical choices.

www.datasociety.net
@datasociety

Illustration: shuoshu@GettyImage