1

PLACING THE CENSUS IN CONTEXT

This chapter provides an introduction to the census in the broadest sense: as a series of datasets, a statistical agency, and a social and political concept. We begin with a summary of the fundamental datasets that are covered in this book and explore how you can use census data in your research with some examples. In doing so, we will touch on concepts that we will cover throughout the text. While this book is primarily a practitioner's guide to working with census data, this chapter provides essential background information so you can better understand and appreciate the importance and value of the census. We will discuss the roles the census plays within American society and how census data fits within the context of the ever-expanding universe of data that includes the federal statistical system, the open data movement, and big data.

1.1 WHAT IS CENSUS DATA?

We can think of the U.S. Census as a collection of datasets about population, housing units, and businesses that is created by the Census Bureau, which is part of the U.S. Department of Commerce. Census data is collected at regular intervals using methodologies such as total counts, sample surveys, and administrative records. After it is collected or generated, census data is summarized to represent counts or estimates of groups of people for different geographic areas. Census geographies, categories, and terminologies are relatively consistent across the different census datasets, and we will explore them in Chapters 3 and 4. A comparative summary of the datasets covered in this book is provided in Table 1.1.

TABLE 1.1 ⬢ COMPARISON OF CENSUS DATASETS COVERED IN THIS BOOK

Dataset	Method	Frequency	Subjects	Geographies	Variables
Decennial Census	100% Count	10 years	Population, housing	Many	Several
American Community Survey	Sample survey 3.5 million addresses	Annual	Population, housing	Many	Many
Population Estimates Program	Administrative records	Annual	Population	Several	Few
Current Population Survey	Sample survey 60k households	Monthly	Population	Few	Many
Business Patterns	Administrative records	Annual	Businesses	Several	Few
Economic Census	100% count and sample survey	5 years	Businesses	Several	Several

When most Americans think of the census, they think of the 10-year or decennial census that is used to gather basic data about the total population. The decennial census is an actual count of people and housing units, and it serves as the baseline for measuring and generating other census datasets. Demographers refer to data that is collected from total counts as enumerations, or simply as populations. The American Community Survey (ACS) and the Current Population Survey (CPS) are ongoing sample surveys of the population that collect detailed demographic and socioeconomic characteristics. Sample surveys collect information from just a small subset of the population, either randomly or from targeted groups, which is used to estimate what the total population is. The size of the sample is carefully determined, so that the sample data can be used to estimate the total population for a given geographic area with a reasonable level of precision. The ACS is a large survey that is published annually for large and small geographic areas, while the CPS is a smaller survey that is published monthly and is summarized for the nation as a whole or for the states. The Population Estimates Program (PEP) is produced from administrative records and other census datasets to create annual estimates for areas like states, counties, and municipalities. The Census Bureau produces data for businesses via the Economic Census, which is a 5-year count of most types of businesses and a sample of other types, and the County and ZIP Code Business Patterns (ZBP), which is created from administrative records on an annual basis.

Who is counted in the census? It varies based on the dataset, and we will cover the specific details about the different methodologies that are used and the variables that are collected in Part II of this book. For now, the short answer is “everyone.” The decennial census counts all people residing in the United States on census day: citizens and permanent residents; documented and undocumented immigrants; people living in households; people living in institutionalized settings like college dormitories, military bases, prisons, and hospitals; and the homeless.

The ACS and CPS are primarily sample surveys of residential addresses, so they do not capture the full spectrum of the population that the decennial census captures. The ACS does sample people living in group quarters (institutionalized settings), but the sample is small enough that it is able to publish coarse estimates only for large areas like states. The PEP is derived from the decennial census and administrative records that include birth and death certificates, so in theory it captures everyone. The business datasets capture most businesses, with some exceptions.

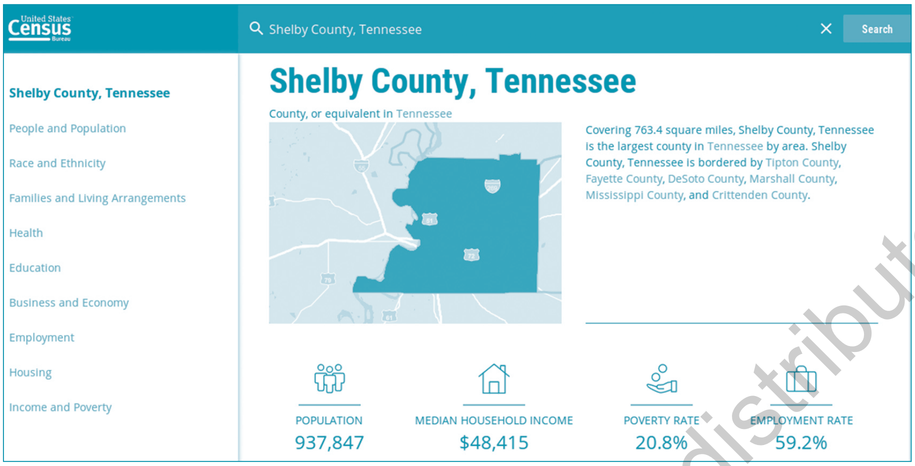
Census data is captured from households, institutions, and businesses through paper and online forms and, when necessary, through on-site visits and canvassing. One of the reasons that the Census Bureau is able to produce reasonably accurate and geographically detailed counts and estimates of the population is that it is a government agency that is backed by law. People are required to fill out and return their census forms. The Census Bureau sends out a series of reminders to nonrespondents, and if a household still does not respond, the Bureau sends an actual enumerator out to interview them for the decennial census and follows up with a sample of nonrespondents in person or on the phone for the ACS. In contrast, private polling agencies would never be able to accomplish a count or survey at the same scale due to the cost of conducting it and their inability to compel people to respond.

1.2 APPLICATIONS OF CENSUS DATA

What can you use census data for? At the simplest level, you may want to look up information for your town, city, or state to get some basic facts to support a story you are writing or research you are doing. The Census Bureau publishes profiles that contain a broad swath of data for one place. With the Bureau’s new data discovery platform, **data.census.gov** (Figure 1.1), a simple place name search will provide you with quick facts, charts, and maps. We will explore this platform in Chapter 2.

Alternatively, you might want to compare one variable for many places in order to see which cities are growing fastest or which areas have the highest income or most unemployment. The Census Bureau publishes comparison tables that you can search through, modify, and download. Or maybe you need to gather many census variables for many places for a research project where you are creating new data, maybe even with data from other sources. The Census Bureau allows you to download data

FIGURE 1.1 ⬢ GEOGRAPHIC PROFILES USING DATA.CENSUS.GOV



Shelby County, Tennessee

Search

Shelby County, Tennessee

County, or equivalent in Tennessee

Covering 763.4 square miles, Shelby County, Tennessee is the largest county in Tennessee by area. Shelby County, Tennessee is bordered by Tipton County, Fayette County, DeSoto County, Marshall County, Mississippi County, and Crittenden County.

- Shelby County, Tennessee
- People and Population
- Race and Ethnicity
- Families and Living Arrangements
- Health
- Education
- Business and Economy
- Employment
- Housing
- Income and Poverty

POPULATION	MEDIAN HOUSEHOLD INCOME	POVERTY RATE	EMPLOYMENT RATE
937,847	$48,415	20.8%	59.2%

in bulk or to access it via a computer program or script using an API (application programming interface).

Or perhaps, you need to visualize census data. You can do this using a number of online tools, or you can download the data and visualize it on your desktop using spreadsheets or geographic information systems (GIS). In this book, we will demonstrate several of these methods to create charts and maps like the examples in Figure 1.2 that depict commuting data for the Northeast Corridor and New York City from the ACS.

As an academic librarian who specializes in geographic datasets, I have helped hundreds of people find, process, and interpret census data for their projects to support arguments in their research and to create new information. This is the kind of data analysis that professor Gary Klass describes in his book *Just Plain Data Analysis* and that we will cover in this text. Klass makes a distinction between "plain data analysis" as processing, presenting, and evaluating statistics to support social and political research as opposed to statistical analysis, which focuses on the testing of hypotheses (Klass, 2012). Here are a few examples that illustrate the kinds of research you can do with census data and what you will learn from reading this book.

- Each semester, I help journalism students with neighborhood reporting projects in New York City. The Census Bureau does not collect data for "neighborhoods" like Midtown, Chelsea, or Harlem, as these are areas that are defined locally. I assist them with translating the Census Bureau's geographies

FIGURE 1.2 ⬢ VISUALIZING CENSUS DATA

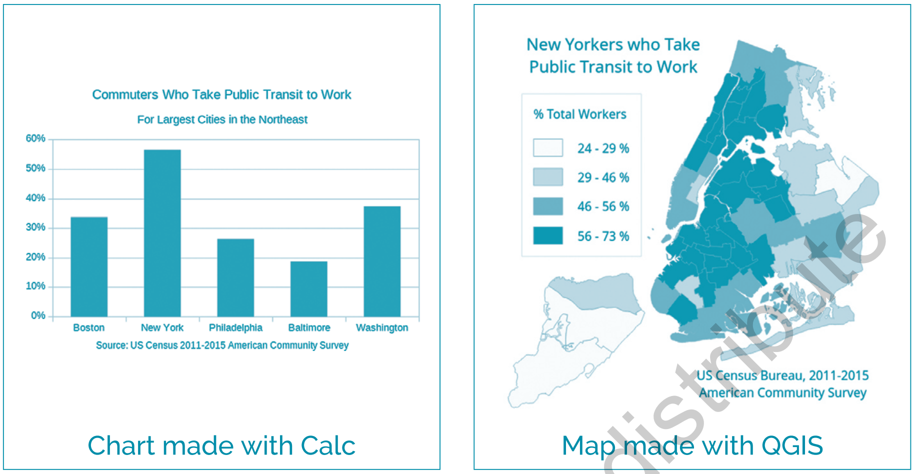


Chart made with Calc

Map made with QGIS

like census tracts or Public Use Microdata Areas into what we consider to be neighborhoods, and we walk through a number of different online sources for census data where they can get profiles. We will cover sources for data in Chapter 2, geography in Chapter 3, creating aggregates for neighborhoods in Chapter 6, and integrating census data into writing in Chapter 9.

- I worked with two journalism professors to combine ACS data and presidential election results in order to identify counties that had low median income, high unemployment, and high poverty compared with the U.S. average and that had switched to voting for the Republicans in 2016 after having voted for the Democrats in the previous two presidential elections. I gathered and loaded the variables into a database, so that we could select counties that met the criteria. The professors combined the results of this analysis with other data to select a county that would serve as a field trip destination for an investigative reporting class. We will cover population groups in Chapter 4 and the ACS in Chapter 6 and will introduce databases in Chapter 5.
- A PhD student was doing research on heat waves, heat-related death, and poverty. She was working with county-level data from many sources from the 1970s to the early 2010s. We not only dove into the historical census files to obtain population and poverty data but also discovered another important variable she could use: From 1960 to 1980, the census asked people whether they had air conditioning in their homes. She was able to use this data for older decades and then created estimates for recent decades using data from the Department of Energy. We will explore historical census data in Chapter 12.

- Our lab advised the New York City Comptroller's Office in creating a series of statistical profiles on the economy of each of the city's Community Districts (New York City Comptroller's Office, 2017). It took data from the ZBP and summed it to ZIP Code Tabulation Areas so the data could be related and assigned to the districts. They collated the ZBP, ACS, and decennial census data into a concise and attractive report and web-based interface. We will cover census geography in Chapter 3, business datasets in Chapter 8, and creating derivatives and relating different geographies in Chapter 11.
- As part of a workshop I teach on GIS, I incorporate an example where we use demographic data from the census, TIGER geographic boundary files, and other geographic data such as the location of subway stations and coffee shops to identify possible locations for opening a new neighborhood coffee shop. We will cover GIS and the Bureau's geographic products in Chapter 10.

1.3 ROLE OF THE CENSUS IN AMERICAN SOCIETY

In this section, we will consider how census data and the Census Bureau fit within the context of American society. In doing so, we will also touch on various aspects of the Census Bureau's history. For a fuller historical treatment, Margo Anderson's *The American Census: A Social History* (2015) is a definitive account, and the history portion of the Census Bureau's website at `https://www.census.gov/history/` is quite comprehensive.

The census has played a vital role in American democracy since the country's founding. The United States was the first country to institute a population census for the purpose of assigning representatives to a democratically elected legislature (Emigh, Riley, & Ahmed, 2016a). Article I, Section 2, of the U.S. Constitution provides the original, legal basis for the census:

> Representatives and direct Taxes shall be apportioned among the several states which may be included in this Union, according to their respective Numbers. . . . The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.

The first census was taken as stipulated in 1790 and has been conducted every 10 years since. The decennial census is used to reapportion seats in Congress based on the differential growth in population between the states, and the data is subsequently used to redraw legislative districts in states that either gained or lost seats. Other provisions in the Constitution provide justification for a federal statistical system. Article I, Section 9, requires that federal appropriations have legal authorization and that the government provides regular statements of its accounts, and Article II, Section 3, stipulates that the president must give Congress an annual update on the state of the union. Statistics were seen as one method for fulfilling these obligations and measuring the nation's progress (Anderson, 2010).

Outside the Constitution, a large body of federal law requires that specific census variables be collected. The statutory uses of each variable that will be collected in the 2020 census and the future iterations of the ACS are published in a report that the Census Bureau submitted to Congress (U.S. Census Bureau, 2017h). For example, the Civil Rights and Voting Rights Acts require data on age, sex, race, employment, and disabilities to evaluate whether civil rights are being protected. Census data is used for allocating hundreds of billions of dollars of funding for federal assistance programs to states and local governments, such as Medicaid, Highway Planning and Construction, Title I Education grants, Temporary Aid for Needy Families, Community Development Block Grants, and more (U.S. Government Accountability Office, 2009). In fiscal year 2016, approximately 320 federal programs used census data to allocate more than $880 billion in federal funds, primarily to state and local governments but also to individuals through direct assistance programs (Reamer, 2017, 2018).

The original decennial census was conducted by U.S. Marshalls, who fanned out across the country on horseback and counted people based on instructions from Congress. As the country grew in size and population and demands for census data increased, the mechanisms for collecting, tabulating, and presenting data grew in complexity and sophistication to meet the demands. Beginning in 1850, a temporary Census Office was established prior to each census to direct operations and tabulate the results (Anderson, 2015, pp. 41–58), and by 1880, this office, staffed with professional statisticians, took control over all census operations (Anderson, 2015, pp. 89–101). In 1902, Congress established the Census Bureau as a permanent office under the Department of Commerce and Labor that remained in operation year-round, and it became one of the chief statistical agencies within the expanding federal government. The number of questions on the census grew from the mid-19th to early 20th century at the instigation of stakeholder groups that included professional statistical societies and business interests. From the mid- to late 20th century, the needs of the federal government for allocating funding and directing policy became the driving force behind the addition and standardization of questions.

Placed squarely in the middle of America's political mechanisms, the census is a strongly debated and contested issue. In their two-volume, comparative, historical study of censuses in the United Kingdom, the United States, and Italy, Emigh et al. (2016a, 2016b) conclude that this intense interaction around the U.S. Census ensures that it remains a vibrant social institution, whereas in other countries, population counting is seen as either a bureaucratic or symbolic exercise since it is disconnected from political outcomes. In the United States, there are fierce debates and lawsuits over how the census is conducted, what questions are asked, how categories for race and ethnicity are defined, and whether the census is accurate or not. Undercounting specific areas or population groups can result in the loss of federal aid and political power for these places or groups.

There are two peculiarities of American government vis-à-vis the census that help ensure that it remains relevant to society at large. First, since it is stipulated as part of the Constitution, it is reasonably assured that the census will be conducted every 10 years in a manner that's relatively consistent. Many other countries have abandoned doing actual counts of the population in favor of using annual sample surveys, estimates based on administrative data, or population registers (Baffour, King, & Valente, 2013). Following years of controversy and lawsuits against the Census Bureau concerning population undercounts in the 1980 and 1990 census, the Supreme Court ruled that the decennial census must be an actual count: It cannot be based on a sample and the count cannot be adjusted using statistical means in any way (Anderson, 2015, pp. 228–247). The Census Bureau can create and adopt new innovations like the ACS, but it must continue to do an actual 10-year count for apportionment purposes.

Second, according to U.S. copyright law (Title 17, Section 105), all works produced by the U.S. government, with few exceptions, are automatically released into the public domain and can be used by anyone for any purpose. This makes the census widely available and accessible, and throughout history, public, private, academic, and nonprofit agencies have employed census data for their own purposes. Stakeholder and interest groups throughout society lobby for changes in the census to meet their needs and also lobby on behalf of the Census Bureau to keep programs funded. In contrast, many other countries copyright their census data and limit what is available. With few exceptions, the United States is rather unique for having a large and established ecosystem of census data users.

Given the accessibility and quality of the census, it is employed for many purposes throughout American society. The Economics and Statistics Administration published a study on the value of the ACS with the subtitle: *Smart Government, Competitive Businesses, and Informed Citizens* (2015) that contains detailed statistics and vignettes on how census data is used. Examples include Kroger supermarkets creating sales projections and siting new stores and academic researchers and governments in Hawaii creating disaster management plans to cope with volcanic eruptions and lava flows.

State and local governments have always been heavy users of census data, because they can use it to study their own communities and create appropriate policies for urban and regional planning, social assistance, public services, and economic development. Within academia, the census serves as a foundational dataset within the social sciences. Compared with many other datasets, the census is geographically detailed, rich in its breadth of variables, relatively consistent, longitudinal, relatively accurate, and well documented. In academic research, the census is used to provide context and a frame of reference for describing places and population groups, can be used as criteria for selecting areas of study, can help define sampling strategies, and serves as the basis for new and derived estimates (Dickason, 2012). For example, census data is particularly important in the field of public health, where it is used as the basis for studying different populations in relation to risk and exposure to public health threats (Wilson et al., 2017).

In the private sector, there are two types of business that routinely use census data: (1) those who use it to make business decisions and (2) information brokers who use it as a commodity. Census data is used extensively within the fields of marketing and real estate. Marketers use it to identify populations and areas that would be good targets for their products and services, while real estate agents use census data to supplement their own information in order to understand housing markets and characterize neighborhoods. Information brokers gather data from many different sources, aggregate it, and use it to produce intelligence that they can sell to third parties, while others create web-based products that can be used for doing research.

Census data plays a key role within American representative democracy for apportioning political power and the resources of the U.S. government. Over time, it has become a vital piece of the nation's infrastructure that is similar to other public goods and services, in that it provides a piece of the foundation on which the country's society and economy rest through the basic yet essential information it provides. Based on the definition of public goods as described by political philosopher Angela Kallhoff in her book *Why Democracy Needs Public Goods* (2011), census data fits the definition as it is nonrival (each individual can use it without affecting someone else's use) and nonexcludable (it is free for all to share). Census data helps generate a public sphere by providing reliable information that creates mutual awareness of others in our society, and it serves as a focal point for debate over issues of common concern.

1.4 CRITICISM OF THE CENSUS

The census is not without flaws or critics. In this section, we'll summarize some of the philosophical and political objections to the census (we will cover issues related to methodology such as undercounting and sample sizes as we discuss each dataset in later chapters). The first and certainly the most earnest concern is the confidentiality and privacy of an individual's responses to the census questionnaires. Throughout the 20th century, federal law stipulated that the Census Bureau would not publish or share records of individual responses to census questionnaires (U.S. Census Bureau, 2009). The current law, established in the 1950s and amended in the 1970s, prohibits the disclosure of individual-level census information for 72 years from the date the census was conducted. Despite these laws, there is a general and growing suspicion of government surveillance (fanned by controversies such as the National Security Agency's PRISM system) and government data-gathering programs from all sides of the political spectrum. Census confidentiality statues were rescinded during the First and Second World Wars under emergency security measures (Anderson, 2015; Aratani, 2018), so there are concerns that it could happen again given some future emergency.

While there is reason for concern, it is important to consider the environment in which the Census Bureau operates. The federal government is composed of hundreds of agencies that operate according to their own missions, needs, and interests and that compete for resources. What's in the best interest of one agency may not be in the best interest of another. The Census Bureau's goal is to create the most accurate population statistics that it possibly can. To achieve this, it must establish a high level of trust with the American people and ensure that each individual's responses will be held in confidence according to the law. Therefore, it is not in the Census Bureau's best interest to share information with other government agencies as it will erode the public's trust and jeopardize the accuracy of the statistics, if people refuse to respond out of fear for their privacy.

In Margo Anderson's (2015) account of the Census Bureau's history, she describes how the Bureau fought to maintain its independence within the federal statistical system so that it could fulfill its mission of generating accurate statistics. In particular, the Bureau successfully resisted every attempt to tie its statistical-gathering activities to other branches of the government that specialized in regulatory enforcement, so that it could reassure individuals and businesses that their data would be used only for the purpose of generating summary statistics.

The Bureau continues with this struggle today. In late 2017 and early 2018, the Justice Department lobbied the Census Bureau to include a question on citizenship on the 2020 census form, which they deemed necessary for upholding the Voting Rights Act and fighting voter fraud (Baumgaertner, 2018; H. L. Wang, 2018h). Given the bitter partisan debates over immigration and the uncertainty and fear among immigrant groups (both legal and undocumented) about their status, the Census Bureau and its supporters (including the two previous secretaries of Commerce under the Obama and Bush administrations; Pritzker & Gutierrez, 2018) strenuously objected to this suggestion. All residents, regardless of their status, are counted in the decennial census. Given deepening suspicion of the government's motives, it's likely that many would refuse to participate and thus would jeopardize the accuracy of the count and all the programs that depend on it. In June 2019 the Supreme Court ruled against the addition of a citizenship question to the 2020 census.

Some members of Congress have suggested that since the census is used for apportioning seats in Congress, either people who can legally vote or only U.S. citizens should be counted. This would be contrary to the intentions of the Founding Fathers and the 220-year history of the census, which has always counted every single person as it was deemed to be the simplest and fairest method for conducting the count. Children cannot vote, but there are approximately 74 million children in the United States and they depend on basic government services like schools. Legal permanent residents cannot vote and are not citizens, but they pay taxes and contribute to society. Politicians are elected to represent all members of their districts, and the Supreme Court agrees. The Court reconsidered the practice of counting every person as opposed to counting eligible voters during the drafting of the Fourteenth Amendment after the Civil War and decided to uphold the population count as the simplest and fairest approach. Since then, the Court has upheld this opinion on several occasions, most recently in 2015, when they ruled that states may count all residents, regardless of whether they are eligible to vote, when drawing legislative districts (Liptak, 2016).

Beyond the issue of confidentiality is a simpler issue of personal privacy that can be summarized as follows: "Why is the government asking me all these questions? It's none of their business!" In this view, the Constitution says that there must be a 10-year count and says nothing about asking other questions or running additional surveys like the ACS. Therefore, some believe that most of the questions are unconstitutional and people have a right to refuse to answer them. However, as discussed in the previous section, there are several sections of the Constitution that provide a basis for establishing a federal statistical system. There are also federal laws and court decisions that require the government to collect statistics in order to fulfill many obligations. The Census Bureau cannot ask questions simply because they might be novel or interesting; every single question is asked because it has some basis in the law.

The Census Bureau explicitly ties each question to the law that requires it and presents this information to Congress (U.S. Census Bureau, 2017h).

In terms of privacy, the image of the government as a 1984 Big Brother that's gathering information about every citizen through coercion seems less plausible given life in the early 21st century. Every day, millions of Americans freely share information about themselves (knowingly and unknowingly) on social media and the internet that is infinitely more personal and potentially compromising than anything they share on a census form. This information is held by technology companies, credit agencies, and data brokers, many of whom sell it to third parties. By and large, these companies are completely unaccountable, and we cannot even know or request what data has been collected about us (Kitchin, 2014). Concerns about the Census Bureau's collection of basic demographic information seems minor in comparison.

The economics of the census is another issue that's frequently raised by fiscal conservatives. The census has been criticized as a waste of tax payer dollars, and it has been suggested that the private sector could do a better job. In reality, the federal statistical agencies' share of federal budget resources represented about 0.04% of gross domestic product in 2016 (Executive Office of the President, 2017), a trifling amount compared with the budgets for defense, Social Security, and Medicare. The private sector cannot compel people to fill out census forms and could not possibly conduct a count or survey of the same scope and detail. Businesses rely on census data the same way they rely on other public goods, such as roads, mass transit, and schools, as fundamental pieces of infrastructure that the economy is built on. When census programs are threatened by budget cuts, business leaders and trade groups are among the first to lobby against them. For example, when the American Community Survey was threatened with cuts in 2012, the Target Corporation collaborated with the Census Bureau to produce a YouTube video that showcased how Target uses census data to tailor its stores and products to different markets (U.S. Census Bureau, 2012b). Many fiscal critics fail to measure the cost of the census against the benefits that it provides (Wilson et al., 2017).

There are good reasons to scrutinize the census. It is important to debate the census questions and categories to ensure that they reflect the changing nature, interests, and needs of our society. It is necessary to highlight issues with methodology that could result in unforeseen consequences regarding the accuracy of statistics. Given the creeping amount of surveillance in our society and the growing number of data breaches, confidentiality must be of utmost concern. But like every other political or public policy issue, it is important to study the underlying arguments to determine whether they are rooted in facts or opinions, either informed or uninformed. The ability to have reliable information for the purpose of checking facts is one of the reasons why we create census data to begin with.

1.5 THE CENSUS WITHIN THE DATA UNIVERSE

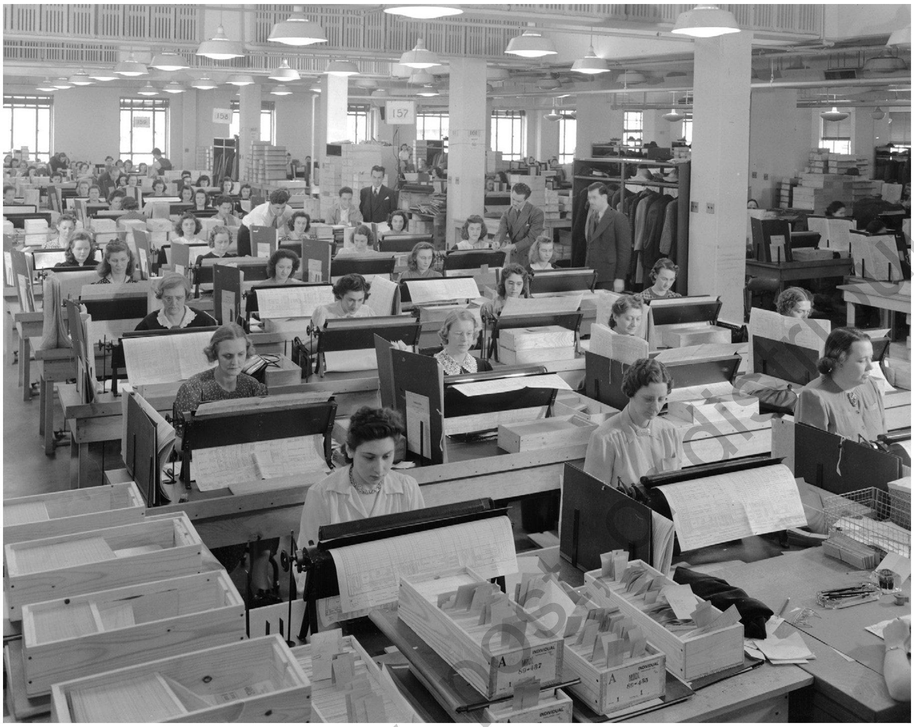
Where does the census fit into our data-saturated world? In this section, we'll situate census data within this context. The census datasets exist as part of a larger federal system of data collecting and publishing activities. The census can also be considered as part of the growing open data movement with some caveats, while it is largely distinct and separate from what most people think of as big data.

The Federal Statistical System

The census is part of the U.S. Federal Statistical System, whose mission is to provide evidence-building functions, which the government describes as "the collection, compilation, processing, analysis, and dissemination of data to create general purpose, policy- and program-specific, and research-oriented statistics and datasets. They also include program evaluation, performance measurement, and public health surveillance" (Executive Office of the President, 2017). The Census Bureau is one of the 13 principal statistical agencies whose primary mission is the production and analysis of statistics, and it receives the largest share of the statistical program's budget ($1.4 billion out of $7.2 billion in 2016).

Given the Census Bureau's size and the depth and breadth of its knowledge for creating statistics, it supports many other federal and state agencies in gathering and creating data. It has a long history of innovation in this field. In the late 19th century, it pioneered the use of mechanical punch card technology for tabulating data (Figure 1.3 shows women reading entries from 1940 census enumerator forms to create punch cards, which would be fed into machines to tabulate results). The Census Bureau was in the forefront of developing statistical sampling methods in the 1930s, which were envisioned as efficient ways for collecting timely data on an ongoing basis (Anderson, 2015, pp. 176–179). Sample survey methods and the Bureau's early adoption of digital computer technology in the mid-20th century allowed for the expansion and growth of data collection and tabulation. In the late 20th century, the Census Bureau helped spread the adoption of GIS in the United States through the creation and distribution of TIGER, a database of geographic boundary files. They were also one of the earliest agencies to publish data on the internet.

FIGURE 1.3 ⬢ CARD PUNCH OPERATORS CREATING POPULATION CARDS FOR THE 1940 CENSUS



Source: National Archives `https://catalog.archives.gov/id/7741405`

Many of the statistical agencies specialize in providing data that is specific to their departments and missions. The Bureau of Transportation Statistics focuses on transportation-related data while the National Center for Education Statistics focuses on education-specific data. The Census Bureau is unique in that its datasets appeal to a broad range of fields and interests; there are questions on commuting, education, the labor force, disabilities, and housing. It is also distinct in its ability to provide detailed data for small geographies that is uniform and comparative; almost all of the other datasets published by the government are not tabulated below the state or county level (Hartnett, Sevetson, & Forte, 2016).

Datasets can be classified into three categories based on how they are created: (1) statistical, (2) administrative, and (3) derived. Most of the Census Bureau's datasets are statistical datasets, while the majority of the other agencies produce data from administrative sources. Statistical datasets are created for the specific purpose of having data to answer specific questions. Statistical datasets can be generated from a total count, like the decennial census, or from sample surveys, like the ACS. In contrast, administrative datasets are created as a by-product of some process. For example, the primary function of the (IRS) is to collect taxes to raise revenue for the federal government, and it uses forms like the 1040 to gather this information. The purpose is to collect taxes, not to produce data. As a by-product, the IRS creates datasets that are used to measure migration between states and counties, based on whether a person's address changed from one year to the next. Derived datasets are data created from other data. The Census Population Estimates Program is a derived dataset that uses data from the decennial census, the ACS, the IRS, the Medicare Enrollment Program, and the National Center for Health Statistics to estimate the annual population for states, counties, and metropolitan areas.

Open Data

Over the past decade, there has been increasing interest around the concept of open data. In the most basic sense, data is considered open if it's free to use, reuse, and redistribute with minimal requirements. The open data movement seeks to build collaboration and participation around free datasets that can be used to study and improve public services and to spur economic growth (Goldtstein & Dyson, 2013). Open data should meet a number of technical requirements that are intended to ensure that it is as accessible as possible. Data should be complete, primary (not summarized), timely, well-documented, and machine readable in nonproprietary data formats, so it can be easily processed without restrictions or expensive tools (Kitchin, 2014).

In many ways, the census could be considered as the original open dataset, as it has been well-documented, widely distributed, and publicly available since its inception in 1790. It falls within the public domain and can be used by anyone for any purpose. It is highly accessible as it can be discovered via several different web interfaces. The data is stored in machine-readable formats (CSV, text-delimited, spreadsheets, database tables, XML), which are formats that have a suitable organization and structure that allows data to be directly retrieved and manipulated. Given the number of datasets and their size and complexity, search engines cannot always crawl and index them directly, but in many instances, users can get persistent URLs to tables so that data can be linked to and cited. It is well-structured and indexable; every record represents a piece of American geography, and that geography is assigned a unique ID code (called a GEOID) that is relatively consistent within and across datasets and years. Census data stored in separate tables can be related and tied together using these identifiers.

One of the challenges for both description and accessibility is the sheer size and complexity of the datasets, which makes them confusing for new and even seasoned users to understand and navigate. The Census Bureau invests a lot of effort in providing tools to cater to many users, and the process for creating datasets is transparent. Each of the Bureau's data discovery tools includes links to glossaries with definitions and terminology, and each of the individual statistical program websites (the decennial census, the ACS, the Economic Census, etc.) includes detailed and frequently updated information that describes the methodology used for collecting and processing the data.

The census is not "complete" in the sense that it's not a primary or secondary dataset. Primary data is data that's collected by an individual or organization for their own use, secondary data is primary data that's distributed for use by outside researchers for their own purposes, and tertiary data is derived data: data that's been aggregated and summarized from the primary set. The Census Bureau's primary data, records of individual responses to census questionnaires, is subject to confidentiality regulations that protect individual's privacy. An individual's responses to the census are kept confidential for 72 years before they can be released, and until then, the data cannot be shared with anyone, including other branches of the government. The Census Bureau provides samples of individual responses with personal identifying information removed in public use microdata files, but not the complete datasets. Most of the data is summarized by population groups and geographic areas, some of which are quite small in size.

Whether the census is timely is a matter of opinion. It is more timely than it used to be, as the detailed socioeconomic characteristics of the population are provided annually as part of the ACS, rather than just every 10 years with the decennial census. The Population Estimates Program and the County Business Patterns are published on an annual basis, and the Current Population Survey is published monthly. But in the big data world where data is provided in real time, the census is not considered timely. It is published at set intervals, and there is a time lag from the time the data is collected to the time it is processed and released.

Given their transparency, accessibility, documentation, structure, and geographic detail, the census datasets do serve as foundational layers in the open data universe. From the open data perspective, they cannot be considered primary or timely, and on these points, we can contrast the census with "big" datasets.

Big Data

In the colloquial sense, the census is quite large, but in the technical sense, it is not big data. Big data is captured in real time and has a granular level of detail, representing a specific person or event at a specific geographic location. Cameras and sensors that constantly monitor the environment are capturing big data, as are websites monitoring clicks, social media sites registering every comment and post, and online forms like 311 requests that are capturing individual complaints. In his book *The Data Revolution*, Rob Kitchin (2014) contrasts big data with small data, and he characterizes the latter as traditional datasets that are produced in a tightly controlled manner with limited scope, size, and time frame. Census data is a prime example of a small dataset: The size and scope are limited to a specific number of questions: if the data is sample based, it is limited to a certain number of respondents, and the time frame ranges from 1 to 10 years. Because of confidentiality reasons, the data is often coarse, summarized by groups and by places. The design is also inflexible; once a count or survey begins, the methods cannot be changed without compromising the dataset and generating great expense.

In contrast, big data seeks to be exhaustive and finely detailed, and is flexible and scalable in production. Kitchin describes big data as being high in velocity (the speed in which it is produced) and volume (the sheer amount that is produced). The allure of big data is the notion that we can simply collect as much information as possible and analyze it in the hope of uncovering trends and making connections and predictions that were previously impossible to conceive without access to modern resources like machine learning and infinite disk space. So why bother with small data like the census when we have big data?

Big data captures what's easy to capture and whatever is openly expressed. It is often represented at face value by technology enthusiasts, even though the data is often not designed to answer specific research questions it's being applied to and is often dirty or unprocessed. While limited in volume and velocity, small data has a long development history with established practices and a design that seeks to answer specific research questions. Kitchin uses the analogy of gold mining; small data studies look for gold in narrow seams while big data studies attempt to extract nuggets from large-scale open pit mining. The principal difference is in investing resources to collect data to answer specific, targeted questions versus searching through tons of big data and hoping it tells us something (Kitchin, 2014). Ultimately, big data has the same limitations as small data; it is merely a representation of reality that is influenced and biased by the context in which it's created. Despite the hype generated around big data, it is not the objective, exhaustive, perfect, and sole solution for answering all of the world's questions.

Let's look at two recent examples that illustrate the limitations of big data versus the value of census data. In 2017, Facebook was touting its strengths to advertisers and investors by saying that its social media platform had the ability to reach 41 million adults in the United States between the ages of 18 and 24, and 60 million adults between the ages of 25 and 34. This sounded pretty impressive, until an analyst at an investment firm checked Facebook's numbers against the latest census data and found that there are only 31 million adults in the United States aged 18 to 24 and 35 million aged 25 to 34 (Hem, 2017; Swant, 2017). When presented with these discrepancies, Facebook responded that their data was designed to estimate how many people in a given area are eligible to see an advertisement that a business might run. Their estimates were derived based on Facebook user behavior, user demographics, and location data from devices and were not designed to match population or census estimates. They concluded by saying that they are always working to improve their estimates.

There are three lessons we can draw from this example. First, big data suffers from the same limitations as small data. The Facebook data was never cleaned or processed to estimate the actual population; it was simply taken at face value and accepted as is. Their statistics can be inflated because people misrepresent their age, have multiple or fake accounts, and because the location services capture people who are in a given area but don't necessarily live there, such as tourists. The census suffers from these same issues; people can be undercounted or overcounted, and there are challenges determining what a person's residence is (people with vacation homes, military personnel overseas, the homeless). The Census Bureau has an advantage relative to Facebook regarding data accuracy, as a person has more to gain by having multiple accounts or lying about their age on a social media platform versus a government form that most people fill out once every 10 years. More important, the Census Bureau has experience with addressing these issues and has methodologies for coping with missing and possibly false information.

This leads to the second lesson. Because the census is in the public domain, the process is transparent. We can go on the census website and freely access the data and all the documentation that's associated with collecting, processing, and disseminating it. People who disagree with the data can lobby the Census Bureau and Congress to try to force changes. While this might be difficult to achieve, you still have the right to do it and have the information at your disposal to create meaningful arguments. In sharp contrast, the Facebook data is in a black box. We can only guess and make assumptions about how it is created. While tech companies constantly push and pull us to freely give them data about ourselves, they resist any attempt to share the data they collect about us with us, or even disclose what they do with it. Many of the big datasets, especially the data generated from social media, suffer from this lack of transparency.

Last, this example illustrates one of the important use cases of census data: The census can serve as a baseline that we can check other datasets against. It can be used for fact checking, as the analyst used it in checking Facebook's claims, and for benchmarking, calibrating, and adjusting population estimates that are generated from other sources, which is what Facebook failed to do.

Meanwhile, *The Washington Post* reported in 2017 that scientists were now able to estimate what the demographic characteristics of different neighborhoods were based on the cars that are parked in the neighborhood (Ingraham, 2017). The researchers collected Google Street View images from 200 U.S. cities, created a schema that correlated the makes and models of thousands of cars with cars in the images, and used this data to build a model that predicts the race, income, and voting characteristics of the population in small census areas. The researchers compared their findings with the ACS data and found a high correlation between their estimates and the actual census data. They suggested that their method could be used to provide more data that is just as accurate as the ACS but could do so in a timely fashion at a fraction of the cost (Gebru et al., 2017).

There are a number of lessons that can be drawn here. Like many experiments that take place in Silicon Valley, the results are novel and interesting, but the exercise takes place in a moral vacuum. Instead of asking a person to identify themselves on a government census form with information that describes how the data will be used, a private company takes pictures of a neighborhood and a third party uses this information to estimate who lives there. Is it just or fair to estimate what a neighborhood's population is like by photographing the cars parked on its streets? The ACS is used to allocate federal funds for everything from transportation projects, to programs for schools, to assistance for needy families. Would it be ethical to use the car-based data to allocate this money, instead of the ACS? Or what if another third party uses this data instead of the ACS to make decisions on whom to give a home loan? While the researchers never explicitly claim that their method should be used to replace the ACS, they implicitly point in this direction as they emphasize how expensive and untimely the ACS data is.

As part of its mission, one of the Census Bureau's goals is to ensure that everyone in the United States is counted as part of the decennial census and that a representative sample of the entire population is included in all their sample-based products. Furthermore, the categories that are used for tabulating the data must represent the entire U.S. population, and the census data itself must be accessible to everyone as a public good.

This mission cannot be fulfilled by the private interests in any one of the examples just described. Researchers in the car study state that they were unable to reliably estimate the presence of children or people employed as farmers, as children don't drive cars and the study omitted rural areas. In the Facebook example, even though a large percentage of the population uses Facebook, there are groups of people that tend to use social media less than others. In the United States, about 7 in 10 Americans used social media in 2018, and people who were older, lower income, or living in rural areas used it less than people who were younger, higher income, and urban (Pew Research Center, 2018). In essence, the big datasets that seek to be exhaustive are not truly exhaustive, but suffer from selection bias based on their context and, in some cases, by conscious decisions made by the people who shape how the data is created.

This does not mean that big data should be dismissed entirely, but it should not be considered as a holy grail. "Small" datasets like the census continue to play a valuable role as high-quality open datasets that are designed to answer targeted questions regarding the demographic and socioeconomic characteristics of the United States. As a public good, the census is transparent, accessible, representative of the entire population, and accountable to the public in ways that private or proprietary datasets cannot be.

1.6 CONCLUSION AND NEXT STEPS

This chapter was designed to give you an overview of the census, so you can understand its legal justification, see the various roles it plays in U.S. society, and place it within the context of a broader data universe. The rest of this book is devoted to teaching you the practical concerns of understanding, finding, retrieving, processing, analyzing, and interpreting census data. As we address these concerns and learn about the different geographies, subjects, and datasets, we will touch on some of the contextual and ethical issues that we covered in this chapter. While many of our concerns will seem practical (How do I represent these racial categories? How can I combine these areas to study a neighborhood? What threshold should I use for establishing some criteria?), decisions made in creating and using data will always have social, political, economic, or ethical consequences.

In the next chapter, we'll get moving right away: We'll go directly to the main source for census data and start exploring the different datasets and tables, and then we'll step back in Chapters 3 and 4 to understand how this data is summarized and organized geographically and categorically.