Personal Privacy and the Public Good

Balancing Data Privacy and Data Utility

Claire McKay Bowen
August 2021
ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people’s lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.
Contents

Acknowledgments iv

Personal Privacy and the Public Good 1
  Tension between Privacy and the Public Good 1
  Record Linkage Attacks 3
  Examples of Record Linkage Attacks 3
  Examples of Data Being Used for the Public Good 5

2020 Census and Data Privacy 6

Methods of Data Privacy and Confidentiality 8
  Remove Personally Identifiable Information 9
  Suppression 9
  Rounding 10
  Adding Noise 10
  Top- and Bottom-Coding 11
  Generalization 12

Measuring Data Utility 15
  Summary Statistics 16

Outcome-Specific Analyses 17

Global Utility Metrics 18

Balancing Privacy and Utility 18
  Defining Traditional Disclosure Risk 19
  Defining Differential Privacy 22
  Executing the Privacy and Utility Workflow 27

Takeaways 28

Notes 29

References 31

About the Author 32

Statement of Independence 33
Acknowledgments

This report was funded by the Tableau Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the author and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute’s funding principles is available at urban.org/fundingprinciples.

The author thanks the following individuals who generously provided invaluable feedback that greatly improved this guide:

- Constance Citro, Senior Scholar, Committee on National Statistics at the National Academies of Science, Engineering, and Medicine
- Diana Elliot, Principal Research Associate, Urban Institute
- Kathy Pettit, Principal Research Associate, Urban Institute
- Izzy Youngs, Research Specialist, Massive Data Institute at Georgetown University
Personal Privacy and the Public Good

The Census Bureau greatly delayed the release of 2020 census data products because of the COVID-19 pandemic and related stay-at-home orders. The 2020 Census was also affected by natural disasters, new household dynamics, eviction freezes, displacement of college students, and national social upheaval. Further, the Census Bureau implemented for the first time a new privacy definition, called differential privacy, for their disclosure avoidance system (that is, its approach for protecting their publicly released data). Leaders from states, counties, cities, and towns rely on this data for school planning, budgeting, social program provision, and redistricting. They are questioning and scrutinizing the quality of the 2020 Census products because of the data equity challenges that have arisen from privacy protection methods, such as ones that use differential privacy, and the impact of enumeration during the pandemic. Data users must understand how important these issues are to have an informed engagement as the Census Bureau considers applying differential privacy to other key datasets, such as the American Community Survey.

In this report, we aim to help readers better understand the tension between privacy and utility (or usefulness) and the challenges of balancing the two sides.

Tension between Privacy and the Public Good

The data and technology landscape has drastically changed in recent decades, motivating the Census Bureau to review and update its privacy protection methodologies for its data products. But many who use these data do not know there is a tension between data privacy and data utility or what challenges the Census Bureau faces. We will start with a simple example using smartphone data.

Imagine that we were to collect every United States resident’s smartphone location data. The data contain basic demographic details (such as age group), the date, time, and latitude and longitude coordinates. All personally identifiable information, such as names, is removed.

Suppose we select one person and observe where he or she is located over time, and we observe the person within the circled location in figure 1 from 10 p.m. to 7 a.m. Most would think this person lives here, because he or she is in a residential area during sleeping hours.

Suppose also that the person is located within the circled area in figure 2 from 8 a.m. to 5 p.m. Most would think the person works in this area, especially if this trend were to continue throughout the work week.
FIGURE 1
Map of Residential Area North of Silver Spring, MD

Source: Map generated using Leaflet with latitude at 39.007216 and longitude at -77.029646.

FIGURE 2
Map of Workplace Area in Washington, DC

Source: Map generated using Leaflet with latitude at 38.884472 and longitude at -77.023873.
With a few pieces of information, we quickly infer where this person lives and works. If we continue this analysis, we can then study this person's personal routine, such as what time he or she goes to work and comes home as well as where this person likes to eat and typically shops.

**Record Linkage Attacks**

We can combine other data to gain additional information about a person. This type of attack is referred to as a record linkage attack, where a malicious actor tries to identify individuals in anonymized data by combining one or more external data sources, such as other public databases or proprietary data. Conducting a record linkage attack becomes easier if the dataset contains unique records. This is because a bad actor can use the unique records' information as identifiers to learn more about that record.

As an example, Sweeney (2000) conducted an analysis on only the 1990 Census data to determine the number of unique observations within certain geographies. She discovered that out of the 248 million total observations,

- 87 percent (215.8 million) could be identified as unique observations based on the five-digit zip code, gender, and date of birth at the national level;
- 53 percent (131.4 million) could be identified as unique observations based on location (i.e., city, town, or municipality), gender, and date of birth; and
- 18 percent (44.6 million) could be identified as unique observations based on gender and date of birth at the county level.

More recently, in 2019, the Census Bureau conducted a reconstruction attack, a method for trying to reconstruct confidential data based on the publicly released data, on the 2010 Census. Census Bureau researchers discovered they could reidentify one-sixth of the United States population using publicly available data, such as from Facebook (Leclerc 2019).

**Examples of Record Linkage Attacks**

The following are more real-life examples of record linkage attacks using other publicly released data outside of the census, suggesting that privacy concerns around decennial census activities are of growing importance.
NETFLIX PRIZE
In 2006, Netflix offered $1 million to improve their show and movie recommendation system by 10 percent and provided access to a dataset containing over 100 million movie ratings from almost 500,000 Netflix subscribers. Although other researchers developed new predictive algorithms, Narayanan and Shmatikov (2008) reidentified Netflix subscribers by using IMDb data. This breach in privacy (which allowed, for example, the prediction of people’s sexuality by their watch history) led to a lawsuit and the cancellation of a follow-up competition.

AOL SEARCHER NO. 4417749
Also in 2006, AOL released an anonymized database of over 20 million web search queries for academic research. Although AOL removed the names and other personally identifiable information, the New York Times still identified one of the AOL users based on her searches such as "landscapers in Lilburn, Ga" and "homes sold in shadow lake subdivision gwinnett county georgia." Once identified, her other searches revealed more personal information that included "60 single men" and "dog urinates on everything."

HEART PROBLEMS AND FIREARMS
Stanford researchers found they could infer additional information about participants in smartphone metadata, such as whether someone has a heart problem or owns an AR semiautomatic rifle (Mayer, Mutchler, and Mitchell 2016).

REIDENTIFICATION OF AMERICANS FROM HEALTH DATA
In 2019, computer scientists from the Imperial College London and Université Catholique de Louvain estimated that they correctly reidentified 99.98 percent of Americans from anonymized health data with 15 attributes including zip code, date of birth, gender, and number of children (Rocher, Hendrickx, and De Montjoye 2019).

REIDENTIFICATION OF BEN BROILI
Knowing when an anonymized person deviates from his or her normal routine can be revealing as well. In 2019, the New York Times covered how much information is gathered from our smartphone location data. The data contained the same information as our earlier example, and the Times team correctly identified Ben Broili, who worked at Microsoft before switching to Amazon, because he changed his work routine.
Examples of Data Being Used for the Public Good

Given these possibilities and examples, should any data (both census and other data) be made publicly available? The risks may be too high. However, the data could be used for the benefit of society.

For instance, the decennial census serves an immediate public good. The federal government uses the data for apportioning the 435 seats for the United States House of Representatives, redistricting state voting lines, planning for natural disasters, allocating the annual $1.5 trillion budget, and more. Local cities and municipalities also use census data to distribute funds for health, education, housing, and infrastructure. Further, nonprofits and other research institutions often combine census data with other datasets to further their research.

The following are other real-world examples of data being used for the public and social good.

CONTACT TRACING APPS
The Centers for Disease Control and Prevention defines contact tracing as the "process of notifying contacts of exposure, addressing questions and concerns, referring for SARS-CoV-2 testing, encouraging self-quarantine, monitoring of symptoms, and assessing the need for additional supportive services during the quarantine period (14 days from last exposure)." With political support and proper integration into public health systems, research shows that contact tracing apps help reduce the spread of COVID-19.

DISASTER RESPONSE
Federal and state governments improve emergency management scenarios by predicting, preparing for, and preventing natural disasters, such as hurricanes, forest fires, and earthquakes. Moreover, emergency planners use location data to develop better evacuation plans and help rescue survivors.

These real-world examples demonstrate the lack of invaluable data that could benefit society.

HOUSING EVICTION
According to the Eviction Lab, "little is known about the prevalence, causes, and consequences of housing insecurity." This is because many cities lack timely eviction data. However, once the data are collected, public policymakers must be careful about how they use the data or they risk worsening the eviction crisis.
PERSONALIZED MEDICINE
A blog post from the National Institutes of Health\textsuperscript{17} poses the question, “Wouldn’t it be nice if treatments and preventive care could be designed just for you?” Personalized medicine is an area of research that uses predictive models to provide specific medical diagnosis, prevention, and more for an individual patient. But access to confidential data is a major challenge for personalized medicine (Brothers and Rothstein 2015).

VETERAN SUICIDE PREVENTION
The Department of Veterans Affairs\textsuperscript{18} prioritizes addressing veteran suicide. However, the Government Accountability Office reported that the department needs to “improve its process to accurately identify all on-campus Veteran deaths by suicide by ensuring that it uses updated information and corroborates information with VA facility officials” (GAO 2020).

This is the tension between balancing personal privacy and the public good. Revealing too much information places people at risk, such as by empowering stalkers and other malicious actors. But collecting too little information restricts our ability to help people, such as with contact tracing apps.

2020 Census and Data Privacy
How does the tension between privacy and utility relate to the 2020 Census? We will walk through another example.\textsuperscript{19}

Imagine the Census Bureau is trying to count someone for the 2020 Census. Suppose this person is a Black millennial woman. If she lived in Washington, DC, the city’s size and racial diversity means that the Census Bureau can easily “hide” the millennial woman in the data while also preserving certain statistical qualities: that there are already many millennial Black women in Washington, DC.

But what if she lived in another area, such as Wyoming, which has a similar population but a different demographic breakdown? As shown in figure 3, “hiding” the Black woman in Wyoming, the least populated state in the United States, becomes much harder. In general, there are very few Black people living in Wyoming as compared with Washington, DC.
What are the options to keep a person's identity hidden when he or she is in a remote area? Data privacy and confidentiality methods (i.e., methods that anonymize data to protect confidential information) can generally protect a person's information (such as his or her race) in two ways:

1. Change the overall demographic representation significantly
2. Remove the specific population from the data

This example shows how population counts in small towns (and small populations generally) can easily be distorted by data privacy and confidentiality methods (e.g., distorting the demographics in Wyoming). Alternatively, privacy can be compromised to maintain accuracy (e.g., leaving the data as is, which could allow the identification of Black people in Wyoming). Data cannot perfectly preserve both privacy and utility: Privacy protections must be lowered to achieve higher utility and vice versa. In the case of the Black millennial woman in a town in Wyoming, if she did not care that people knew she came from a small town, the Census Bureau could lower privacy barriers to provide more accurate statistics.

Why then doesn't the Census Bureau lower its privacy protections? Many might not care if their demographic information is fully exposed and would want an accurate census. But in addition to the record linkage attacks we discussed previously, other situations might arise where the knowledge that people with certain demographics live in an area could legally and ethically violate privacy.
The Census Bureau is bound under Title 13 of the US Code\textsuperscript{20} to “provide strong protection for the information [that the Census] collect[s] from individuals and businesses.” For this reason, the Census Bureau must carefully balance the need to protect every person and business in the United States while preserving data accuracy. Moreover, considering the legacy of internment camps during World War II, many Asian Americans (specifically those of Japanese descent) might not be comfortable with public data revealing that Asian Americans live in a particular small town.\textsuperscript{21}

Methods of Data Privacy and Confidentiality

In this section, we will cover several methods of privacy and confidentiality, which statisticians refer to as statistical disclosure control or limitation, that the Census Bureau implements. Although Census Bureau data can be accessed in other ways, such as through secure enclaves called Federal Statistical Research Data Centers, the focus of this report is on data being anonymized for public release. The anonymization includes releasing aggregates and microlevel (or record-level) data.

To help explain the statistical disclosure control methods, suppose we are working with the local government to release microlevel socioeconomic data of those who live in Washington, DC (table 1). These data would be valuable to help the local government better target their economic stimulus programs. However, we are concerned about sharing this data publicly given the data have demographic and income information. We decide to apply several statistical disclosure control methods. For more details on how the Census Bureau applies statistical disclosure control methods, see work by Lauger, Wisniewski, and McKenna (2014).

\begin{table}[h]
\centering
\caption{Fictitious Washington, DC, Socioeconomic Data} \\
\textit{A fictitious socioeconomic dataset with participants’ names, ages, education levels, and income}
\begin{tabular}{lccr}
\hline
Name    & Age  & Education & Income  \\
\hline
Alex     & 28   & Bachelors & $51,489  \\
Andrea   & 26   & Bachelors & $36,072  \\
Bob      & 92   & Some college & $0      \\
Beth Ann & 58   & Doctorate & $77,226  \\
Daniel   & 17   & High school & $623    \\
Donna    & 32   & Bachelors & $41,543  \\
Edward   & 45   & Bachelors & $115,879  \\
Elizabeth & 53  & Masters & $99,253  \\
\hline
\end{tabular}
\end{table}

\textbf{Source:} Author.
Remove Personally Identifiable Information

We first want to remove any personally identifiable information or other variables that we do not want to include in the public version of the data. We can either replace them with nondescript IDs or remove the variables entirely. As seen in table 2, we replace the name values with numbers.

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Education</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>28</td>
<td>Bachelors</td>
<td>$51,489</td>
</tr>
<tr>
<td>02</td>
<td>26</td>
<td>Bachelors</td>
<td>$36,072</td>
</tr>
<tr>
<td>03</td>
<td>92</td>
<td>Some college</td>
<td>$0</td>
</tr>
<tr>
<td>04</td>
<td>58</td>
<td>Doctorate</td>
<td>$77,226</td>
</tr>
<tr>
<td>05</td>
<td>17</td>
<td>High school</td>
<td>$623</td>
</tr>
<tr>
<td>06</td>
<td>32</td>
<td>Bachelors</td>
<td>$41,543</td>
</tr>
<tr>
<td>07</td>
<td>45</td>
<td>Bachelors</td>
<td>$115,879</td>
</tr>
<tr>
<td>08</td>
<td>53</td>
<td>Masters</td>
<td>$99,253</td>
</tr>
</tbody>
</table>

Source: Author.

Suppression

After removing the personally identifiable information, we must decide whether any values in our data should be suppressed. Suppression, or not reporting certain values from the data, is one of the earliest and easiest statistical disclosure control methods. For example, if only one African American woman lives in Natrona County in Wyoming, then we would want to suppress or not report that person. The threshold or rules for suppression vary by institution, type of data, and other factors, and these can all be subjective.

For the 2010 Census and the American Community Survey Public Use Microdata Samples, the Census Bureau only released a sample of microlevel data in certain areas if it satisfied their geographic threshold rules. These rules enforce the threshold to be at least 100,000 records, but could be higher depending on the region, detail of the variables, whether the survey was longitudinal (data repeated over time), and whether other similar public data are available. Several federal agencies use the rule of three, meaning values are not reported if there are fewer than three records in any combination of attributes.

Suppose our data has more than 100,000 records, so the Washington, DC, socioeconomic data will not be suppressed. If it had fewer records, it would not be available.
Rounding

The data reports income to the nearest dollar, which is too specific: malicious actors can more easily conduct record linkage attacks with the precise income values. We can round these values in several ways. One statistical disclosure control rounding scheme has a randomization for rounding up or rounding down. This would be done by applying a statistical method in the data to randomly apply rules for which records get rounded and in which direction. For example, if someone reported an income of $987, then this income has a 70 percent chance of being rounded up to $990 or 30 percent chance of being rounded down to $980.

The Census Bureau uses the following for dollar amounts (table 3).

- $0 remains $0
- $1–7 rounded to $4
- $8–$999 rounded to nearest $10
- $1,000–$49,999 rounded to nearest $100
- $50,000+ rounded to nearest $1,000

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Education</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>28</td>
<td>Bachelors</td>
<td>$51,500</td>
</tr>
<tr>
<td>02</td>
<td>26</td>
<td>Bachelors</td>
<td>$36,100</td>
</tr>
<tr>
<td>03</td>
<td>92</td>
<td>Some college</td>
<td>$0</td>
</tr>
<tr>
<td>04</td>
<td>58</td>
<td>Doctorate</td>
<td>$77,000</td>
</tr>
<tr>
<td>05</td>
<td>17</td>
<td>High school</td>
<td>$620</td>
</tr>
<tr>
<td>06</td>
<td>32</td>
<td>Bachelors</td>
<td>$41,500</td>
</tr>
<tr>
<td>07</td>
<td>45</td>
<td>Bachelors</td>
<td>$116,000</td>
</tr>
<tr>
<td>08</td>
<td>53</td>
<td>Masters</td>
<td>$99,000</td>
</tr>
</tbody>
</table>

Source: Author.

Adding Noise

We can also protect the data by adding or subtracting random values to the sensitive information in the record, such as income. The values could be drawn at random within certain bounds (e.g., -5 to 5) or based on a probability distribution. As an example for the latter, we could draw values from a bell curve centered at zero, meaning there is a higher chance of drawing very small values close to zero.
and a lower chance of drawing very high values at the tails of the curve. Privacy researchers refer to this approach as adding noise, infusing noise, sanitizing results, or perturbing the data.

For our socioeconomic data example, we add random noise to the age variable. Table 4 shows the new age values, where the random noise is drawn from a bell curve–shaped distribution (i.e., normal or Gaussian distribution). We see that some of the values added or subtracted are very small (0, 1, and 4) with a few larger values (6 and 8). Adding and subtracting values at random brings in more uncertainty, making it harder for a potential attacker to know what the original age value was.

**TABLE 4**
Infusing Noise
*Dataset with random noise added to the age values*

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Age with Noise</th>
<th>Education</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>28</td>
<td>28-4=24</td>
<td>Bachelors</td>
<td>$51,500</td>
</tr>
<tr>
<td>02</td>
<td>26</td>
<td>26+1=27</td>
<td>Bachelors</td>
<td>$36,100</td>
</tr>
<tr>
<td>03</td>
<td>92</td>
<td>92+8=100</td>
<td>Some college</td>
<td>$0</td>
</tr>
<tr>
<td>04</td>
<td>58</td>
<td>58-6=52</td>
<td>Doctorate</td>
<td>$77,000</td>
</tr>
<tr>
<td>05</td>
<td>17</td>
<td>17+0=17</td>
<td>High school</td>
<td>$620</td>
</tr>
<tr>
<td>06</td>
<td>32</td>
<td>32+1=33</td>
<td>Bachelors</td>
<td>$41,500</td>
</tr>
<tr>
<td>07</td>
<td>45</td>
<td>45+4=49</td>
<td>Bachelors</td>
<td>$116,000</td>
</tr>
<tr>
<td>08</td>
<td>53</td>
<td>53-1=52</td>
<td>Masters</td>
<td>$99,000</td>
</tr>
</tbody>
</table>

*Source: Author.*

**Top- and Bottom-Coding**

Even after adding noise, the example data still has extreme values or outliers for age. Specifically, ID 03 is at the upper end for the age distribution of Americans. Also, suppose for the data, there are very few working individuals who are under the age of 18. Instead of publishing the exact age for ID 03 and 05, we limit how large or small the values are that we report. This restriction is called top- and bottom-coding. We can top- and bottom-code our Age category to substitute “greater than 85” and “less than 18” rather than the exact values (table 5).
TABLE 5
Top- and Bottom-Coding
Dataset with the age values top- and bottom-coded

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Education</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>24</td>
<td>Bachelors</td>
<td>$51,500</td>
</tr>
<tr>
<td>02</td>
<td>27</td>
<td>Bachelors</td>
<td>$36,100</td>
</tr>
<tr>
<td>03</td>
<td>&gt;85</td>
<td>Some college</td>
<td>$0</td>
</tr>
<tr>
<td>04</td>
<td>52</td>
<td>Doctorate</td>
<td>$77,000</td>
</tr>
<tr>
<td>05</td>
<td>&lt;18</td>
<td>High school</td>
<td>$620</td>
</tr>
<tr>
<td>06</td>
<td>33</td>
<td>Bachelors</td>
<td>$41,500</td>
</tr>
<tr>
<td>07</td>
<td>49</td>
<td>Bachelors</td>
<td>$116,000</td>
</tr>
<tr>
<td>08</td>
<td>52</td>
<td>Masters</td>
<td>$99,000</td>
</tr>
</tbody>
</table>

Source: Author.

Generalization

Although having more attributes or levels for variables is ideal for some analyses, the finer-grain data increase the likelihood that unique records can be more easily identified in a record linkage attacks. To increase privacy protection, the Census Bureau requires that all categorical variables have at least 10,000 people nationwide in each published category. The categories that do not satisfy this stipulation have to be placed into broader groups.

For our example, we can generalize education into coarser groups that should reduce or eliminate the number of unique observations. In table 6, we convert the education levels of “high school,” “some college,” “bachelors,” “masters,” and “doctorate” to the broader education levels of “no college,” “bachelors,” and “graduate degree.”

TABLE 6
Generalization
Dataset with the education values generalized to broader levels

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Education</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>24</td>
<td>Bachelors</td>
<td>$51,500</td>
</tr>
<tr>
<td>02</td>
<td>27</td>
<td>Bachelors</td>
<td>$36,100</td>
</tr>
<tr>
<td>03</td>
<td>&gt;85</td>
<td>No college</td>
<td>$0</td>
</tr>
<tr>
<td>04</td>
<td>52</td>
<td>Graduate degree</td>
<td>$77,000</td>
</tr>
<tr>
<td>05</td>
<td>&lt;18</td>
<td>No college</td>
<td>$620</td>
</tr>
<tr>
<td>06</td>
<td>33</td>
<td>Bachelors</td>
<td>$41,500</td>
</tr>
<tr>
<td>07</td>
<td>49</td>
<td>Bachelors</td>
<td>$116,000</td>
</tr>
<tr>
<td>08</td>
<td>52</td>
<td>Graduate degree</td>
<td>$99,000</td>
</tr>
</tbody>
</table>

Source: Author.
DATA SWAPPING

The Census Bureau implemented a new disclosure control technique called data swapping for the 1990 Census and continued to use it through the 2010 Census. This method involves swapping the data between households in different locations that had similar variable characteristics. They did not swap all observations; only records with the highest disclosure risks were targeted. Because data swapping applied within a specific geographic area, the approach did not affect the population or other characteristic totals. More importantly, data swapping allowed the data to be released down to the block level, restoring many of the suppressed tables from past censuses.

Data privacy researchers first introduced data swapping for contingency tables or tabular results, but the method could be extended to microlevel data at the cost of a higher swap rate. In other words, going back to the privacy-utility trade-off, if we want more accurate (e.g., microlevel) data, then we have to increase the amount of records we swap to add more uncertainty into the data. Another drawback to data swapping is that it doesn’t preserve multivariate relationships.

For example, suppose our data are subset by certain neighborhoods within Washington, DC. The first 100 records are from Hill East (a neighborhood on the southeastern side of the city), and the next 100 records are in Georgetown (on the northwestern side). In table 7, we decide to swap records 03 and 05 with other records from Georgetown that have similar age and education values because there are very few records with these extreme age values and no college education. Although the new 05 record has a similar income value, record 03 does not; it has changed from $0 to $250,000. The average income value for the Hill East neighborhood has potentially increased by a lot, which could change how economists determine whether a neighborhood needs more financial aid.

TABLE 7
Swapping
Dataset with records 03 and 05 swapped with similar other records with similar age and education values from another Washington, DC, neighborhood

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Education</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>24</td>
<td>Bachelors</td>
<td>$51,500</td>
</tr>
<tr>
<td>02</td>
<td>27</td>
<td>Bachelors</td>
<td>$36,100</td>
</tr>
<tr>
<td>03</td>
<td>&gt;85</td>
<td>No college</td>
<td>$250,000</td>
</tr>
<tr>
<td>04</td>
<td>52</td>
<td>Graduate degree</td>
<td>$77,000</td>
</tr>
<tr>
<td>05</td>
<td>&lt;18</td>
<td>No college</td>
<td>$1,800</td>
</tr>
<tr>
<td>06</td>
<td>33</td>
<td>Bachelors</td>
<td>$41,500</td>
</tr>
<tr>
<td>07</td>
<td>49</td>
<td>Bachelors</td>
<td>$116,000</td>
</tr>
<tr>
<td>08</td>
<td>52</td>
<td>Graduate degree</td>
<td>$99,000</td>
</tr>
</tbody>
</table>

Source: Author.
As we saw with this example, the impact of data swapping depends on what characteristics were swapped. This means some subpopulations of the data could be unintentionally affected more, such as in our example with low- versus high-income neighborhoods. Further, data swapping is not a transparent technique: swap rate cannot be reported publicly because malicious actors could reverse-engineer the method.

**SAMPLING**

A common statistical disclosure control approach for protecting microdata files that are released to the public, in addition to such methods as top-coding, is to select a subsample of respondents. The Census Bureau pioneered public-use microdata sample files when it released a 1-in-1000 sample of respondents from the 1960 census.

Statistical agencies use subsampling to introduce some form of “plausible deniability” to protect data in public-use microdata files produced from censuses and sample surveys. The general idea is that if someone tried to identify a record in one released dataset with another publicly available dataset, they cannot guarantee the match is correct because the released data are a random subset of the original.

As an example, suppose a bad actor tries to identify someone in their mid-thirties with a bachelor’s degree. The bad actor notices a record within our previous fictitious data that matches this description; ID 06. However, the person the bad actor is trying to identify could plausibly deny the record is theirs, because the released data are a random sample of the entire population of Washington, DC, which already has several college graduates in their mid-thirties.

**SYNTHETIC DATA**

In recent decades, synthetic data has become one of the most popular statistical disclosure control methods among privacy researchers. In May 2021, the Census Bureau announced it will make the 2025 American Community Survey data fully synthetic. This data product is extensively used for analyses on “poverty, inequality, immigration, internal migration, ethnicity, disability, transportation, fertility, marriage, occupations, education, and family structure.”

Synthetic data consist of pseudo or “fake” records that are statistically representative of the original, confidential data. Imagine we collect information on where people traveled for a conference in Washington, DC. The confidential data show that 50 out of the 100 participants are already from Washington, DC. One way to generate synthetic data for this sample is to flip a coin 100 times and report the number of heads results as the number of people from Washington, DC.
Statisticians originally developed synthetic data to address missing data in clinical trial scenarios. Patients often drop out of such studies because they last for several months or years. The statisticians created new observations or values for the missing data by developing a model based on the remaining patient data. The idea of synthetic data is attractive to federal agencies because they contain only "fake" records. But most federal agencies don't use synthetic data yet. This is mostly because of limited human resources (i.e., lack of practitioners who are knowledgeable of the methods and can implement them) and computational resources (i.e., code and proper computing equipment).

In general, synthetic data can be created either based on a model or not based on a model (i.e., with parametric or nonparametric methods). At a high level, the non-model-based approaches calculate the estimates or percentages of counts from the data and use those estimates as weights for a weighted random sampling scheme. Our earlier Washington, DC, conference example would be considered a non-model-based approach, where the weight for the randomization scheme is 50 percent.

Model-based methods rely on estimating or learning an appropriate model based on the confidential data; “fake” records are then created from the model. As an example, suppose we collect the heights of our conference attendees. When we plot the data, we see that the distribution of attendees' heights is similar to a bell curve or normal distribution and decide to use that model to generate our synthetic data.

The use of synthetic data relies heavily on selecting an appropriate model to preserve the data's statistical features, and this reliance has a few potential drawbacks. One is that the analyst must be careful when selecting and using a model that perfectly replicates the confidential data. Some privacy researchers advise splitting the data into multiple parts so that one part can help inform and develop the model while other parts help verify the model’s quality. Another concern is that if an analyst selects a poor model, the synthetic data will provide improper results for data users. This also means that developing a model to capture every interesting feature in more complex data without recreating the confidential data is extremely difficult.

**Measuring Data Utility**

How does the Census Bureau next measure the usefulness of the anonymized data? In general, those responsible for the confidential data should establish data-quality metrics based on how other researchers, institutions, and government agencies will use the data. However, determining which
specific metrics to implement is itself an entire research field. It is impossible to predict all possible analyses that data users might implement and ensure the data will provide valid results for each of those analyses.

Here we review a few utility measures, but many other types exist. Table 8 provides example education data on the education level of people in the Washington, DC, area.

**TABLE 8**

*Example Washington, DC, Education Data*

*A fictitious dataset with the percent breakdown for each education level*

<table>
<thead>
<tr>
<th>Education</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>9%</td>
</tr>
<tr>
<td>High school or GED</td>
<td>17%</td>
</tr>
<tr>
<td>Some college</td>
<td>13%</td>
</tr>
<tr>
<td>Associate</td>
<td>3%</td>
</tr>
<tr>
<td>Bachelors</td>
<td>25%</td>
</tr>
<tr>
<td>Masters</td>
<td>21%</td>
</tr>
<tr>
<td>Professional</td>
<td>8%</td>
</tr>
<tr>
<td>Doctorate</td>
<td>4%</td>
</tr>
</tbody>
</table>

Source: Author.

**Summary Statistics**

Many privacy researchers will first examine the summary statistics of the anonymized data as a quick and easy starting point for assessing utility. Some common summary statistics measure how well the released data preserve the counts, means, and correlations for each variable or combination of variables. The researcher then reports the distance between the original and noisy results. Typically, these measures are bias, root mean squared error, or other distance measures.

For the 2020 Census, the Census Bureau announced the following utility measures and how it defined them: mean absolute error, mean numeric error, root mean squared error, mean absolute percent error, coefficient of variation, total absolute error of shares, and percent difference thresholds equaling the count of absolute percent difference above a certain threshold. Moreover, these measures are often averaged across geographies. For example, mean absolute error is the "average absolute value of the count difference for a particular statistic" and is calculated as the absolute value of the difference between the anonymized data and the confidential data results.
Table 9 shows the mean absolute error between the confidential and the anonymized education data from table 8. We can sum the mean absolute error from table 9 and report a total mean absolute error of 14 percent. We could also report the average of the mean absolute error across groups to equal 1.75 percent. However, researchers must be cautious about how they report the summary statistics or interpret the results. Some summary statistics do not capture the data distribution, where there could be sizable errors within particular groups or geographies.

For instance, our mean absolute error in table 9 ranges from 1 to 3 percent, which is very small. Imagine the mean absolute values are 0, 3, 5, 0, 1, 4, 0, and 1 instead. The average mean absolute error is still 1.75 percent, but we now have values that range from 0 to 5 percent instead of 1 to 3 percent.

**TABLE 9**

**Summary Statistics: Mean Absolute Error**

*A fictitious education dataset with the percent breakdown for Washington, DC, with the anonymized data percent and the mean absolute error*

<table>
<thead>
<tr>
<th>Education</th>
<th>Confidential data percent</th>
<th>Anonymized data percent</th>
<th>Mean absolute error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>9%</td>
<td>10%</td>
<td>1%</td>
</tr>
<tr>
<td>High school or GED</td>
<td>17%</td>
<td>15%</td>
<td>2%</td>
</tr>
<tr>
<td>Some college</td>
<td>13%</td>
<td>11%</td>
<td>2%</td>
</tr>
<tr>
<td>Associate</td>
<td>3%</td>
<td>2%</td>
<td>1%</td>
</tr>
<tr>
<td>Bachelors</td>
<td>25%</td>
<td>27%</td>
<td>2%</td>
</tr>
<tr>
<td>Masters</td>
<td>21%</td>
<td>18%</td>
<td>3%</td>
</tr>
<tr>
<td>Professional</td>
<td>8%</td>
<td>7%</td>
<td>1%</td>
</tr>
<tr>
<td>Doctorate</td>
<td>4%</td>
<td>6%</td>
<td>2%</td>
</tr>
</tbody>
</table>

*Source: Author.*

**Outcome-Specific Analyses**

Privacy researchers may and should ask the data users what analyses they typically implement as another measure for data quality. The privacy community refers to this type of utility measure as an outcome-specific metric. The Federal Register Notice, the Committee on National Statistics Demonstration Products Workshop, and other outreach groups suggested additional metrics to evaluate the 2020 Census. One suggestion was to measure the distribution of federal funds by county, by incorporated place, and by minor civil divisions (e.g., parishes in Louisiana). The Census Bureau announced it will assess the quality for this use case using the total absolute error of shares metric by county within each state as a share of that state, by incorporated place as a share of that state, and by minor civil divisions as a share of that state.
Global Utility Metrics

Using global utility metrics or discriminant-based algorithms is another way to evaluate data quality that is becoming more popular in literature but is not used by the Census Bureau to date. The global utility metrics attempt to measure how close or similar the anonymized data are to the confidential data. A simple example would be comparing two anonymized datasets, one with 10 percent of records that have “less than high school” education and one with 13 percent, where the former is closer to the confidential data value of 9 percent. Most global utility metrics compare against several variables rather than one or two.

Global utility metrics first combine the confidential data with the publicly released data and mark each record as being from the confidential data or the anonymized, public data. Next, the privacy researcher must decide what classification model to use to discern whether a record is from the confidential or the anonymized data. If the classification model "struggles" to assign a record to either the confidential or public data, privacy researchers assume that the two datasets are similar. More specifically, each record receives a probability of being classified as being from the confidential data or the public data. A probability close to 50 percent means the classification model cannot predict any better than a coin flip.

Finally, depending on the global utility measure, the method distills those probabilities into a single value or multiple values to convey how similar the released data are to the original data. The "accuracy" depends on what classification model privacy experts use, because each classification model will measure different characteristics of the data. Privacy researchers need to conduct more scientific studies to fully understand these differences. They also need to explore how to make these methods more computationally efficient for the average data user.

Balancing Privacy and Utility

How do the Census Bureau and other institutions balance the competing needs of privacy and utility for their public data releases? Most expect a one-size-fits-all approach in striking this balance. But these institutions must weigh several factors, such as current laws, expected use cases, and anticipated privacy threats.

In general, privacy researchers follow this workflow when trying to strike the right balance between privacy and utility:
1. Determine the threshold of acceptable disclosure risk, disclosure risk measurements, utility metrics, and invariants.

2. Remove personally identifiable information and variables that are too sensitive to include in the released data.

3. Select the variables in the confidential data that need to be anonymized, which are usually all remaining variables.

4. Develop and apply the statistical disclosure control methods (such as those we described in the previous section) that will reduce the specific disclosure risk measurements and will preserve the chosen utility metrics.

5. Compare the data quality of the released data to the confidential data using the utility metrics.

6. Repeat Steps 4 and 5 if the disclosure risks are too high or if the data utility results are too low.

We have not covered how to define disclosure risk (step 1 of the workflow). There are currently two broad categories of defining disclosure risk: the traditional definition and the differential privacy definition. The former is more intuitive but ad hoc because it involves coming up with what the privacy expert and others think is the privacy threat to the data, whereas the latter is confusing but provides strong privacy protection. We will discuss both broadly and use another fictitious dataset to help explain them.

Defining Traditional Disclosure Risk

As mentioned, the traditional privacy definition is more intuitive because the privacy researcher must state how a malicious actor will attack the confidential data. One example can be found in work by Sweeney (2000), who uses the five-digit zip code, gender, and date of birth to identify 87 percent of the observations in the 1990 Census data. If we expect someone might attack the 1990 Census data in this manner, we would want to reduce the number of records that have unique combinations before releasing that data publicly. Our disclosure risk measure could then be, "How many unique records are present in the data?" If we have a high number of unique records, then we have a higher probability that an intruder can determine who or what a particular record is.

Another way a bad actor may attack the data is by inferring trends in the data. Suppose someone develops a model that predicts a person’s race based on a publicly available dataset that contains first name, last name, gender, age, and zip code. That model could be sold to a mortgage lending company
that can discriminate against applicants without directly asking for applicants' race. In this situation, privacy researchers might want to suppress data or apply other statistical disclosure control methods to limit the amount of demographic and geographic information in the dataset.

As an example, imagine a confidential dataset containing the number of Health Care and Social Assistance, Accommodation and Food Service, and Educational Services establishments at the census-tract level within a county (table 10). We follow the three general disclosure limitation methods that the Bureau of Labor Statistics implements on their Quarterly Census of Employment and Wages data product. These data are counts of establishments, employees, and wages reported by employers.

**TABLE 10**
**Example Industry Establishment Data within a County**
*A fictitious dataset with the number of different establishments in each census tract within a county*

<table>
<thead>
<tr>
<th>Tract</th>
<th>Health care and social assistance</th>
<th>Accommodation and food services</th>
<th>Educational services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tract 1</td>
<td>6</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Tract 2</td>
<td>15</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Tract 3</td>
<td>3</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Tract 4</td>
<td>4</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>County total</td>
<td>28</td>
<td>34</td>
<td>15</td>
</tr>
</tbody>
</table>

*Source: Author.*

**CELL DOMINANCE**

Cell dominance requires that any cells that have less than three contributors (such as people or households in the data) are suppressed automatically. This is another example of the rule of three. In table 11, we suppress all cells with less than three industry establishment contributors.

**TABLE 11**
**Cell Dominance**
*A fictitious industry establishment dataset with values less than three suppressed*

<table>
<thead>
<tr>
<th>Tract</th>
<th>Health care and social assistance</th>
<th>Accommodation and food services</th>
<th>Educational services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tract 1</td>
<td>6</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Tract 2</td>
<td>15</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Tract 3</td>
<td>3</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Tract 4</td>
<td>4</td>
<td>6</td>
<td>--</td>
</tr>
<tr>
<td>County total</td>
<td>28</td>
<td>34</td>
<td>15</td>
</tr>
</tbody>
</table>

*Source: Author.*
**P-PERCENT RULE**

The p-percent rule determines whether a cell should be suppressed based on the contributions to that cell. If after excluding the top two contributors the remaining sum is less than a certain percentage of the top contributor’s value, the cell is not suppressed. The idea is that the rule protects the top contributor from the second contributor and vice versa (Confidentiality and Data Access Committee 2005).

For example, let $p$ be 50 percent. The Health Care and Social Assistance values have the two largest contributors as 15 and 6. The sum of the remaining values is 7, which is less than 7.5 (i.e., $15 \times 0.5 = 7.5$). This means that Health Care and Social Assistance must be suppressed. If we repeat this for the remaining columns in table 12, we do not suppress any more values.

**TABLE 12**

**P-Percent Rule**

*A fictitious industry establishment dataset with the health care and social assistant values suppressed based on the p-percent rule of 50 percent*

<table>
<thead>
<tr>
<th>Tract</th>
<th>Health care and social assistance</th>
<th>Accommodation and food services</th>
<th>Educational services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tract 1</td>
<td>--</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Tract 2</td>
<td>--</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Tract 3</td>
<td>--</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Tract 4</td>
<td>--</td>
<td>6</td>
<td>--</td>
</tr>
<tr>
<td>County total</td>
<td>28</td>
<td>34</td>
<td>15</td>
</tr>
</tbody>
</table>

*Source: Author.*

**SECONDARY DISCLOSURE**

Secondary disclosure requires that any grouping of records cannot have only one contribution suppressed either by the cell dominance or p-percent rule steps. In table 11, for instance, we removed the Education Service value for tract 4 because of the cell dominance. If we report the Educational Services values as is, a bad actor could calculate the suppressed tract 4 value by calculating $15 - 3 - 5 - 5 = 2$. Following secondary disclosure rules, we randomly suppress one of the values remaining in the Educational Services column in table 13.

Overall, the general data privacy community views these traditional definitions as easy to understand and explain. But these intuitive definitions come at a cost: They heavily rely on an accurate understanding of how a malicious actor will attack the data and sometimes lead to more information than desired being removed.
Many attack vectors are possible. Will the attackers target one person, or a group of people? Is the target a specific person, or anyone with specific characteristics? Does the attacker have a powerful computer or several external datasets?

This guessing game becomes more complicated when trying to predict what data and technologies will be released in the future. Some of the statistical disclosure control methods were created in the 1940s to 1980s, when computers were nonexistent or primitive. If the privacy researcher incorrectly assumes the attacker’s behavior, then the privacy protection guarantee for the confidential data can be significantly weakened or useless.

The lack of transparency is another drawback of traditional statistical disclosure control definitions and methods. Both data swapping and the $p$-percent rule rely on not reporting the swap rate and the $p$ value, respectively, to avoid bad actors from reverse-engineering the methods. They and other methods rely on "security through obscurity," which reduces trust between data users and those who anonymize the data.

### Defining Differential Privacy

Essentially, differential privacy takes the traditional privacy definitions and throws them out the window. Differential privacy provides a very strong privacy guarantee for the confidential data by not making the same assumptions that traditional definitions make. In other words, differential privacy does not make predictions on how the malicious actor will attack the data or what external knowledge that person might have to disclose more sensitive information. Differential privacy sacrifices an intuitive privacy definition and produces potentially messier data in exchange for greater privacy protection. This situation has caused several ongoing challenges in educating the scientific community.

---

<table>
<thead>
<tr>
<th>Tract</th>
<th>Health care and social assistance</th>
<th>Accommodation and food services</th>
<th>Educational services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tract 1</td>
<td>--</td>
<td>9</td>
<td>--</td>
</tr>
<tr>
<td>Tract 2</td>
<td>--</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Tract 3</td>
<td>--</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Tract 4</td>
<td>--</td>
<td>6</td>
<td>--</td>
</tr>
<tr>
<td>County total</td>
<td>28</td>
<td>34</td>
<td>15</td>
</tr>
</tbody>
</table>

**Source:** Author.

**TABLE 13**

**Secondary Disclosure**

*A fictitious industry establishment dataset where a random value from educational services is suppressed based on secondary suppression rules*
In general, differential privacy is defined through two key points. First, differential privacy is a mathematical definition or condition that a method must satisfy to be differentially private. Basically, differential privacy is a statement about the method rather than the data. Privacy experts will commonly refer to the algorithms as “differentially private algorithms” or “algorithms that satisfy differential privacy.” Second, differential privacy uses the concept of a privacy loss “budget,” often denoted mathematically as $\epsilon$, to help explain the definition in nontechnical terms. The privacy loss budget quantifies the amount of information being leaked with every released statistic or data.

In other words, if privacy researchers spend more of the privacy loss budget (or a larger value of $\epsilon$), they will gain more accurate information about the data when applying their analyses. But this greater accuracy means less privacy is guaranteed because more information is being “leaked.” Inversely, they could spend a smaller amount of the privacy loss budget, resulting in less accurate information being obtained from the data but providing more privacy protection.

Essentially, $\epsilon$ becomes a knob that researchers can adjust to navigate the trade-off between privacy and utility. With $\epsilon$ set to zero, the released data has no relationship to the confidential data: privacy is total, but the released data have no utility. Setting the privacy loss budget to zero is impractical, so it is never seen in literature or real-world applications. As $\epsilon$ approaches infinity, the released information will have no noise and thus no privacy, but utility will be extremely high. Also, the privacy loss budget must be selected prior to releasing the statistic or the published data, but how to set the value and who sets it is still an open question.

The amount of noise or change that a differentially private method applies also depends on how sensitive the statistical analysis is. This sensitivity is not measured in terms of personal or private information but rather in how robust or resistant the information is to being influenced by outliers. Differential privacy quantifies this sensitivity by measuring how much the result of an analysis or an answer to a question changes in the confidential data given the absence or presence of the most extreme possible person that could be in the population but might not be observed in the data. In other words, if we do not know how the bad actor will attack the data, we should assume the worst-case scenario: that the attacker has information on every observation of the data but one and has unlimited computational power. This means differential privacy tries to account for any possible version of the data that could exist, protecting against future data releases and new technologies. If the answer to our question is too sensitive to outliers for any possible version of the data, then we should add more noise to protect the records that are currently present in the data.
To help understand this concept, imagine the data we want to protect contains socioeconomic information. The question we have is, “What is the median wealth?” According to differential privacy, we must consider the change of the most extreme possible record that could exist in any given data that has demographic and financial information. For our example, that person is Jeff Bezos, who was the wealthiest person in the world in 2020. If Bezos is changed or removed in the data, the median wealth should not change too much. This means we can provide a more accurate answer by adding less noise to the median income statistic, because it is less sensitive to outliers such as Bezos. Consider, however, the question, “What is the average wealth?” Unlike the previous statistic, the answer would significantly change if Bezos were changed or removed from the data. To protect the extreme case, a differentially private algorithm would need to provide a significantly less accurate answer by adding more noise.

For a specific example, let us walk through how to apply the most basic differentially private algorithm in literature, the Laplace mechanism. This mechanism adds noise by drawing values from a Laplace distribution, where the distribution is centered at zero and the distribution variability (i.e., how wide or narrow the distribution is) is the ratio of the privacy loss budget, $\epsilon$, over the sensitivity of the target statistics. Having the distribution centered at zero means there is a higher probability of adding very little or no noise to the confidential data statistics. For the noise variability, if $\epsilon$ is large or the sensitivity of the statistic is low, then there is a higher probability of adding very little noise to confidential data statistic. If $\epsilon$ is small or the sensitivity of the statistic is high, then there is a higher probability of adding a lot of noise to the released statistic.

Figure 4 shows the shape of the Laplace distribution under these conditions. The top image shows a Laplace distribution when the privacy loss budget is large or the sensitivity of the statistic is small. This means that we have a high probability of adding noise close to zero and a low probability of adding large values from the distribution tails to our confidential statistic. The bottom image instead shows a Laplace distribution when the privacy loss budget is small or the sensitivity of the statistic is large. This means that we have a high probability of adding large values to our confidential statistic because the Laplace distribution has wider tails.
Laplace Mechanism

The Laplace distribution centered at zero with the distribution variability is the ratio of the privacy loss budget and the sensitivity of the statistic.

![Diagram of Laplace distribution]

To further demonstrate this, we will use our industry establishment data from earlier and add Laplace noise to the data. We will assume that the county totals will stay the same or are invariant. This means that the total counts must be the same as the confidential total counts of 34.

We next need to calculate the sensitivity for count statistics by asking how much a count statistic could change if a person were changed or removed from the data. In other words, if someone is in the data or not in the data, the count statistic would at most change by 1. This is true for any counting statistic. With the sensitivity calculated, the Laplace distribution variability will be dependent on the privacy loss budget (figure 4).

Table 14 shows the added Laplace noise with $\epsilon = 0.5$ and table 15 shows the added Laplace noise with $\epsilon = 1$. For both tables, the first column contains the confidential counts, the second column contains the random Laplace noise being added, and the third column readjusts the noisy counts to sum to the invariant county total of 34. Because the privacy loss budget increased from 0.5 to 1, we can imagine the bottom Laplace distribution from figure 4 to be the distribution we draw values for table 14 and the top Laplace distribution for table 15. This means that table 14 will have more noise.
added than table 15. As a quick check, we can calculate the mean absolute error to verify that the data utility should improve with increased privacy loss budget. We find that table 14 has a mean absolute error across the tracts of 9 and table 15 has a mean absolute error of 4.

**TABLE 14**

Laplace Mechanism with \( \epsilon = 0.5 \)

*A fictitious accommodation and food services establishment dataset with added noise*

<table>
<thead>
<tr>
<th>Tract</th>
<th>Accommodation and food services</th>
<th>Accommodation and food services with noise</th>
<th>Accommodation and food services readjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tract 1</td>
<td>9</td>
<td>9 - 0.99 ( \approx 8 )</td>
<td>34 ( \times (8 / 38) \approx 7 )</td>
</tr>
<tr>
<td>Tract 2</td>
<td>11</td>
<td>11 + 6.02 ( \approx 17 )</td>
<td>34 ( \times (17 / 38) \approx 15 )</td>
</tr>
<tr>
<td>Tract 3</td>
<td>8</td>
<td>8 - 1.91 ( \approx 6 )</td>
<td>34 ( \times (6 / 38) \approx 5 )</td>
</tr>
<tr>
<td>Tract 4</td>
<td>6</td>
<td>8 - 1.91 ( \approx 6 )</td>
<td>34 ( \times (8 / 38) \approx 7 )</td>
</tr>
<tr>
<td>County total</td>
<td>34</td>
<td>38</td>
<td>34</td>
</tr>
</tbody>
</table>

*Source: Author.*

**TABLE 15**

Laplace Mechanism with \( \epsilon = 1 \)

*A fictitious accommodation and food services establishment dataset with added noise*

<table>
<thead>
<tr>
<th>Tract</th>
<th>Accommodation and food services</th>
<th>Accommodation and food services with noise</th>
<th>Accommodation and food services readjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tract 1</td>
<td>9</td>
<td>9 + 0.90 ( \approx 10 )</td>
<td>34 ( \times (10 / 36) \approx 9 )</td>
</tr>
<tr>
<td>Tract 2</td>
<td>11</td>
<td>11 + 0.73 ( \approx 12 )</td>
<td>34 ( \times (12 / 36) \approx 11 )</td>
</tr>
<tr>
<td>Tract 3</td>
<td>8</td>
<td>8 - 1.47 ( \approx 6 )</td>
<td>34 ( \times (6 / 36) \approx 6 )</td>
</tr>
<tr>
<td>Tract 4</td>
<td>6</td>
<td>6 + 1.63 ( \approx 8 )</td>
<td>34 ( \times (6 / 36) \approx 8 )</td>
</tr>
<tr>
<td>County total</td>
<td>34</td>
<td>36</td>
<td>34</td>
</tr>
</tbody>
</table>

*Source: Author.*

From our Laplace mechanism example, we easily calculated the worst-case scenario; the most a count can change is 1. We did this calculation because differential privacy requires that the data be protected by assuming any possible individual could be in the data and that a data intruder could possess any currently known or unknown potential future information. This last point deviates significantly from the traditional privacy definitions, which often assume the intruder only has the currently available information. When developing differentially private algorithms, privacy researchers must determine all possible records that could exist within the data (i.e., “the universe of possible datasets”) to guarantee differential privacy’s high standard of protection.

Although tabular statistics are easy to calculate, figuring out the universe of possible records for other data types or statistics can be challenging (e.g., data that do not have an obvious upper limit, such as income data). Further, many privacy experts struggle to grasp the concept of “the entire
universe of possible datasets," which highlights the main drawback of differentially private approaches: they are difficult to understand.

**Executing the Privacy and Utility Workflow**

Suppose we have data that needs to be released publicly with some privacy protection. We'll follow the steps laid out earlier for striking a balance between privacy and utility. For step 1, we must first decide the appropriate disclosure risk (usually chosen by a federal agency's disclosure review board and how they interpret certain rules and laws), what disclosure risk measure to use (traditional or differential privacy), and how to assess the data quality (e.g., summary statistics). To help decide, we should ask ourselves the motivation for developing a data privacy method, who will benefit from data access, and whether those responsible for the data release sought input from the data users when trying to answer the previous two questions. Answering these questions while working with people with a wide range of use cases will help determine a good strategy to protect the data while providing sufficiently accurate statistics.

In step 2 of the workflow, we eliminate variables that are too sensitive to release, which is commonly decided by the individuals or institutions who are responsible for the data. Sometimes privacy researchers can help with the decision by suggesting additional variables to get rid of that may seem harmless but can reveal confidential information indirectly (such as in our mortgage lending example).

Based on the variables removed, we can decide if all the remaining variables should be included when developing our data privacy and confidentiality methods (step 3). Consider which variables are required for essential use cases, desired for secondary use cases, and totally unnecessary for any use case. Assigning variables or other information as high, medium, or low priority will help us determine which parts of the data need to be changed and to what extent.

Steps 4, 5, and 6 (applying the statistical disclosure control methods, comparing data quality, and repeating as necessary) are critical for researchers to fine tune the statistical disclosure control methods and ensure that the acceptable levels of disclosure risk and utility are met. Oftentimes, the process of trying to make these decisions is like "holding sand." Shifting or changing one part of the workflow, such as trying to improve the data quality, can easily result in the disclosure risk "spilling out" in unexpected ways.
For example, a statistical disclosure control method may generalize or aggregate the number of people from census tract up to the county level. The generalization prevents too many unique individuals from being identified based on demographic detail at the smaller geographies. However, a data user needs to know accurate counts of Hispanic people to more equitably distribute federal funds to rural regions of the country. This is why developing a method that balances data privacy and data utility is an art as well as a science.

**Takeaways**

Some people believe that data can be both completely private and provide accurate enough information for any analysis. But as we have learned, this is not the case. The data’s privacy protection erodes with every piece of information released, and the usefulness of the data erodes with every step taken to protect privacy.

We also learned there are a lot of moving parts to consider when trying to balance privacy and utility. Data researchers must ask themselves the following:

- What are the data privacy concerns?
- What are the data users’ needs?
- What data privacy and confidentiality methods should we apply?
- What utilities metrics should we implement to assess data quality?
- What definition of privacy should we use?

When answering these questions, we should consider the various perspectives of those involved in the whole process to ensure equitable data access and representation. By learning and engaging with a wide range of data users, the data community can better debate and weigh in on these questions as more federal agencies consider implementing various privacy preserving methodologies. Data users can then better judge if they understand how the data has been adjusted, if the utility metrics illustrate the quality for their use cases, and if the released data are fit to answer their questions.
Notes

1 For more information about the Netflix Prize, see the contest website at https://www.netflixprize.com/.


10 See the National Hurricane Center and Central Pacific Hurricane Center data archive at https://www.nhc.noaa.gov/data/.


14 See the Eviction Lab website at https://evictionlab.org/.


This draws from an existing Urban Wire blog post to explain the tension between privacy and utility in context of the 2020 census. See Claire McKay Bowen, "Will the Census’s Data Privacy Efforts Erase Rural America?" Urban Wire, March 4, 2020, https://www.urban.org/urban-wire/will-censuss-data-privacy-efforts-erase-rural-america.


The US Census Bureau pioneered the use of probability sampling to reduce cost and burden on the public in the 1930s, asking a handful of questions in the 1940 census on a sample basis. See “History | 1940 Overview,” US Census Bureau, last revised May 12, 2021, https://www.census.gov/history/www/through_the_decades/overview/1940.html.


“Incorporated place is established to provide governmental functions for a concentration of people as opposed to a minor civil division, which generally is created to provide services or administer an area without regard, necessarily, to population. Places always are within a single state or equivalent entity, but may extend across county and county subdivision boundaries. An incorporated place usually is a city, town, village, or borough, but can have other legal descriptions.” Definitions from "Geography Program | Glossary, US Census Bureau, last revised September 16, 2019, https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_14.

Establishments are a single economic unit, such as a mine, farm, factory, or store. Establishments are typically at one physical location and differ from a firm, or a company, which is a business and may consist of one or more establishments.
References


About the Author

Claire McKay Bowen is the lead data scientist for privacy and data security at the Urban Institute. Her research focuses on assessing the quality of differentially private data synthesis methods and science communication. In 2021, the Committee of Presidents of Statistical Societies identified her as an emerging leader in statistics for her technical contributions and leadership to statistics and the field of data privacy and confidentiality.
Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.