Journal of Economic Perspectives—Volume 38, Number 2—Spring 2024—Pages 181–200

Government Data of the People, by the People, for the People: Navigating Citizen Privacy Concerns

Claire McKay Bowen

At 5:00 AM, my Android alarm goes off. When I hit the snooze button, the alarm rings again after ten minutes. To help me wake up, I use the Chrome browser on my phone to access US and New Mexico news articles from sources like the *New York Times* and the *New Mexican*. At 5:30 AM, the Spotify app plays my "Morning Jam" playlist while I head to the local pool. At 6:00 AM, my Garmin watch uses "Indoor Pool Swim" to record my distance, times, speed, and heart rate. When I am done, my watch automatically syncs with the Garmin Connect app, which then updates my other workout apps, TrainingPeaks and Strava. TrainingPeaks allows my triathlon coach to view my workout details, whereas Strava is a social media platform for workout enthusiasts. At 7:35 AM, I use Venmo to pay my swim coach, which records an online payment between two individuals with a note saying, "for June."

I head home and begin my workday. The Outlook app logs workday start and end times, along with details of meetings and participants. Box and Dropbox record changes made to documents, including additions, removals, and updates. GitHub records "pull requests" when my collaborators and I discuss possible changes to some code. Google Scholar tracks my search history. Overleaf logs access times for working on a paper. Slack records messages in various channels.

When my workday ends, I take my two dogs for a run in an area just outside Santa Fe, New Mexico. Garmin, TrainingPeaks, and Strava record an "Outdoor

■ *Claire McKay Bowen is a Senior Fellow, Urban Institute, Washington, DC. Her email address is cbowen@urban.org.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.38.2.181.

Run" with distance, time, speed, and heart rate data. The title on Strava is edited to "EZ Joggo with my Doggos 🐕🐕" for personalization. I capture a photo of our run and share it on Instagram and Facebook, along with pictures of my two cats. I end my day watching the British Bake-off on Netflix, while knitting and using Chart-Minder, a knitting app, to track my progress.

This brief narrative highlights the personal data that private companies collect and record about me on a typical day. Although you do not know other personal details like when and what I eat (although there is an app for that, too), you still get more than a glimpse of my professional and personal life.

Tracking other seemingly innocuous information over time can provide additional disclosure. As one example, I started competing in endurance races, like triathalons, in college. If someone linked my publicly available registration information across multiple race events, they could infer that a change in my last name indicates a marriage or that a new home state implies a move in that particular year.

I give my students an assignment along these lines: For one day, they must document what data are collected about them from the moment they wake up until they go to bed. Based on their data logs, they write an essay addressing whether the data is protected by law, potentially beneficial or harmful uses of the data, and equity and ethical effects of collecting this information.

As my students and I both readily admit, most of us willingly use web browsers, engage with social media platforms, use financial apps, participate in rewards programs, and more. While many of these services are free in monetary terms, these companies use consumer data for profit, such as targeted marketing to specific demographics. As a society, we have (mostly) come to accept this reality, although there are evolving practices and regulations that, to some extent, give people some opportunity to opt out of information sharing.

But the types and amounts of personal information collected keeps expanding. Some universities now require students to install an app that tracks their movements on campus (as reported by Belkin 2020). With this information, professors teaching large classes can know their students' punctuality, tardiness, or class absences. Beyond attendance records, one can imagine some social good from this data collection. In the event of a natural catastrophe or a mass shooting on campus, the tracking app could alert students, identify safe refuges, and notify emergency contacts. On the other side, this tracking can be invasive, especially for students who rarely leave campus except during holidays, because the university can have a comprehensive record of their locations for 24 hours a day.

Government collects personal data in two main ways. One is through surveys like the Current Population Survey and the American Community Survey that can provide local and state leaders detailed demographic data about individuals and households in the United States. The other is through administrative data collected for purposes of administering programs, like income tax, data collected from employers to run unemployment and workers' compensation programs, payments made to households or on behalf of households through unemployment

insurance, Medicaid, and others.[1] Again, tracking the individual-level data from these programs over time would reveal more than just looking at a point in time. Moreover, combining government data with proprietary data from private companies that are not subject to federal data consumer privacy laws could reveal more individual data as well.

This essay will focus on the protection of individual-level government data, with an emphasis on survey and administrative data. I like to say that these data are "of the people, by the people, and for the people." The data are *of the people*, in the sense that people do care about their privacy and their confidential data. Although they may be willing to trade off information a bit at a time to private-sector actors for useful purposes—like my workout and knitting apps—many people would be deeply unhappy if their personal data was widely available.[2] The data are also *by the people*, in the sense that government collection of people's information is supported by taxpayer dollars. Therefore, one could argue that anonymized individual-level data should be accessible to data users—such as data practitioners, external researchers, or public policymakers. Literally volumes of research could be cited here about how increased access to government data results in social good *for the people*. As one example, Chetty, Friedman, and Rockoff (2014) showed how elementary school teacher quality has a substantial effect on economic outcomes later in life—a finding that was only possible when the economists had direct access to administrative data. For another example, Nagaraj and Tranchero (2023) discovered that applied researchers with confidential data access through the Federal Statistical Research Data Centers are more likely to produce papers that are cited in public policy reports.

While the data privacy community agrees that these data should be more widely accessible, what to protect in that data and how to do so are highly and intensely debated. The community involved in this conflict can be usefully divided into four groups (Williams and Bowen 2023): (1) *data users and practitioners* consume data, such as analysts, researchers, planners, and decision-makers; (2) *data privacy experts or researchers* specialize in developing data privacy and confidentiality methods; (3) *data curators, maintainers, or stewards* are responsible for data safekeeping, and in that sense are sometimes said to "own" the data; and (4) *data intruders, attackers, or adversaries* try to gather sensitive information from the confidential data. The discussion that follows will refer to all four groups.

[1]While "administrative data" often refers to what is collected by the government for programmatic purposes, the term can be more broadly used to cover, for example, other organizational records. For example, colleges and universities have data student applications, class registration, dormitory assignments, and grades, and private companies have administrative data for purposes like tracking orders.

[2]This article uses the terms "data privacy" and "date confidentiality" interchangeably, but readers should note that the terms are sometimes used with separate meanings. In some cases, data confidentiality refers to how the data privacy community protects participants' information in the data, such as who should have access to the sensitive data under what restrictions. Data privacy refers to the amount of personal information individuals allow others to access about themselves.

Throughout this paper, my discussion will emphasize the fundamental trade-off between data privacy and data usefulness—and how determining an appropriate balance can be difficult. In extreme cases, a perfectly protected dataset would never be released, which is useless to data users. Conversely, a perfectly useful dataset would be released without any statistical data privacy methods applied or any data protections, which would violate privacy concerns. These extreme cases demonstrate how it is impossible to achieve perfect privacy and utility. This is why, even with statistical methods to protect privacy in any given use of data, some information is inevitably leaked with each release of a dataset or statistic. Another challenge is repeated data or statistic releases can gradually erode the overall privacy protection, ultimately reaching a point where the level of protection becomes equivalent to releasing the data without any alterations. This highlights why data curators collaborate with privacy researchers to find a balance between these two extremes and prevent excessive information disclosure.

Traditional Methods of Accessing Data

Protecting and Releasing Public Data and Statistics

For decades, government agencies produced public use data and statistics. During the pre-computer era in the early to mid-1900s, public use data were available to those who braved the "government documents" section of research libraries or those who physically went to government offices to inspect available files. But government agencies typically reported summary statistics, such as total spending on unemployment insurance and total number of people receiving it. Knowing such information on a county-by-county or metro-area basis did not pose much threat to privacy.

As the computer era arrived, government agencies started to provide more detailed public use data that could be directly accessible to both researchers and the general public. For example, the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) releases a public use file for data users based on administrative taxpayer data. Several organizations, such as the American Enterprise Institute (DeBacker, Evans, and Phillips 2019), the Urban-Brookings Tax Policy Center (McClelland et al. 2019), and the National Bureau of Economic Research (Bierbrauer, Boyer, and Peichl 2021), develop microsimulation models based on this public use file that inform the public on potential impacts of tax policy proposals.

To ensure that these public use data protect individual privacy, extensive statistical data privacy methods are implemented. Here, I will use a fictitious socioeconomic dataset to illustrate a range of such methods. For a more comprehensive overview of these methods, Matthews and Harel (2011) offer a detailed review, while McKenna and Haubach (2019) summarize the specific statistical data privacy methods employed by the US Census Bureau.

Suppose the fictitious micro-level socioeconomic dataset contains hundreds of records for individuals residing in Santa Fe. The top panel of Table 1 displays a

Table 1A

Fictitious Santa Fe, New Mexico, Socioeconomic Data

A fictitious socioeconomic dataset with participants' names, ages, education levels, and income

Name	*Age*	*Education*	*Income*
Peter	63	Master's	$51,214
Patricia	48	Bachelor's	$89,464
Ryan	24	Bachelor's	$27,893
Rachel	58	Bachelor's	$74,770
Steve	17	High school	$623
Suzanne	32	Doctorate	$135,883
Tomas	81	Bachelor's	$0
Tiffany	21	Some college	$11,428

Source: Author.

Table 1B

Adjusting Data to Preserve Anonymity

The fictitious socioeconomic dataset (Table 1A) that has been altered with statistical data privacy methods

ID	*Age*	*Education*	*Income*
01	60	Graduate degree	$51,000
02	55	Bachelor's	$89,000
03	25	Bachelor's	$27,900
04	52	Bachelor's	$75,000
05	19	No college	$620
06	37	Graduate degree	$136,000
07	85	Bachelor's	$0
08	17	No college	$11,400

Source: Author.

sample of eight records from the dataset, which includes the person's name, age, education, and income. The bottom panel shows how these data have been adjusted to preserve anonymity, using a series of steps.

As a first step, most personally identifiable information, such as names, should be removed from the data. An obvious step is to replace names with numbers. Data curators, who are responsible for safeguarding the data, may generate individual-level identification numbers if they plan to link the data with other information. If there is no intention to link the data with another source, the variable may be entirely removed.

After removing the personally identifiable information, the most common statistical data privacy method is suppression, which involves the removal of certain values from the data. This approach is easy and quick to implement. As an example, when I attended high school in a remote area of Idaho, I was the only Asian American student. Even with names removed, a data intruder could identify me in a

dataset that included information on race/ethnicity. To ensure my privacy, such information could be removed or suppressed.

Another privacy concern in the fictitious dataset is the reporting of income values to the nearest dollar. To make the records less identifiable, we can round the income values. Instead of rounding to the nearest hundred or thousand, some rounding methods introduce randomization in rounding up or rounding down significant figures. For instance, consider an individual with an income of $596. If we want to round the value to the closest $10, then there is a 60 percent probability of rounding the income up to $600 and a 40 percent probability of rounding it down to $590. There are also other rounding schemes, such as the one utilized by the US Census Bureau, which we implement for the fictitious dataset in the bottom panel of Table 1. In this approach, $0 is rounded to $0, $1–7 rounded to $4, $8–$999 rounded to nearest $10, $1,000–$49,999 rounded to nearest $100, and $50,000+ rounded to nearest $1,000.

Another statistical data privacy method is known as generalization, aggregation, or categorical thresholding. When applying this method, the detailed information is consolidated into broader categories. In our example, we can generalize the education groups, which would decrease or eliminate the number of distinct observations. The bottom panel of Table 1 demonstrates how we changed the education levels of "high school," "some college," "bachelor's," "master's," and "doctorate" into broader categories such as "no college," "bachelor's," and "graduate degree."

Adding or subtracting random values is another popular statistical data privacy method. One way to generate random values is within specific boundaries (for example, –10 to 10) or based on a probability distribution (for example, a bell curve centered at zero). This method is known as adding noise, injecting noise, sanitizing results, or perturbing the data. In the bottom panel of Table 1, noise has been added to the age variable, resulting in new age values. The random noise is drawn from a bell curve–shaped distribution, such as a normal or Gaussian distribution. We see that some of the added or subtracted values are very small (like 0, 1, and 2), while a few are larger values (for example, 6 and 7). Introducing random values creates some uncertainty, making it more challenging for a malicious actor to discern the original age value.

Accessing Federal Data Directly

Over the years, government agencies have been moving slowly toward allowing more data users direct access to the underlying cleaned data,[3] under strict controls. An example of direct data access is through a secure enclave, such as the Federal Statistical Research Data Centers.[4] This secure enclave became available in 1982

[3]Privacy researchers often distinguish two types of confidential data: the original data and confidential data (Hu and Bowen 2023). The former is the uncleaned, unprotected version of the data, such as the raw census microdata. The latter is cleaned—edited for inaccuracies or inconsistencies—and stripped of some personally identifiable information like names. This essay focuses on the latter.

[4]"Federal Statistical Research Data Centers are partnerships between federal statistical agencies and leading research institutions. FSRDCs provide secure environments supporting qualified researchers

(then called the Center for Economic Studies), after data users demanded access to better quality data when the US Census Bureau became more aggressive with its applications of statistical data privacy methods on its data products.

Although more secure facilities are becoming available (for example, the National Science Foundation Secure Data Access Facility[5]), researchers face several challenges to obtain this direct access. Full access to these data is only available to select government agencies, a limited number of data users working in collaboration with analysts from those agencies, or through highly selective research programs administered by these agencies. Further, data users are often required to be US citizens, undergo lengthy clearance processes to gain direct access (which can take months or years), and submit extensive research proposals.

Another challenge is accessibility to these secure facilities. The 33 Federal Statistical Research Data Centers across the United States may seem like enough to be geographically accessible to most data users. But that is not the case. These data centers are primarily located in places with large academic institutions. For me, living in Santa Fe, New Mexico (the state capital), the closest is a 7.5-hour drive to Boulder, Colorado. Moreover, remote access to these data centers grants access to only a limited selection of confidential data and requires a setup that simulates a secure enclave working station, which may prove challenging for many individuals.

Developing Another Tier of Data Access: Verification and Validation Servers

In recent years, synthetic data generation has become a popular method for producing public data that releases more useful information than the other traditional statistical data privacy methods while protecting privacy. The general concept of synthetic data is generating pseudo or fake records that preserve the structure and statistical relationships in the confidential data. Ever since Rubin (1987) proposed the original method to create multiple imputations for missing data, the research community has developed several different flavors (partially and fully synthetic) and approaches to generate synthetic data (such as Bayesian synthesis models). Hu and Bowen (2023) provide a more detailed review of synthetic data for privacy protection.

There are obvious concerns over whether a new analysis carried out on synthesized data will reflect the underlying data. As an example, the Urban Institute and Statistics of Income Division at the IRS have created a synthetic version of the SOI public use file that matches certain statistical properties of the underlying data, such

using restricted-access data while protecting respondent confidentiality." For details, see https://www.census.gov/about/adrm/fsrdc.html (accessed on January 15, 2024).

[5] The National Science Foundation Secure Access Facility provides authorized researchers secure remote access to National Center for Science and Engineering Statistics data and metadata, such as the Survey of Earned Doctorates and the national Survey of Recent College Graduates (https://www.norc.org/research/projects/nsf-secure-data-access-facility.html, accessed on January 15, 2024).

as means, variances, and covariances, and provides reliable estimates from microsimulation models (Bowen et al. 2022a, 2022b). The drawback is that this synthetic SOI public use file (and any other synthetic data) may not perform well for other types of analyses that were not accounted for when synthesizing the data, and thus may not yield reliable parameter estimates for more complex statistical models.

To address this issue, the data privacy community has proposed the use of verification and validation servers, which we delineate here following Williams et al. (2023). A "verification server" is a system that provides information about the quality of statistical inference derived from publicly released data. In other words, the verification server might report whether inferences about the sign, statistical significance, magnitude of the estimates, or other elements derived from the public data or synthetic data are consistent with the confidential data. As one example, Barrientos et al. (2018) synthesized data and created a pilot verification server for the US Office of Personnel Management, which allows the study of career paths and pay differentials for federal employees.

A "validation server" allows users to submit and run statistical analyses on the confidential data after the users have developed those analyses using the publicly released data. For instance, data users could apply a tax microsimulation model on the synthetic public use file. If the data users have other statistical analyses, they can then develop and debug programs using the synthetic data to ensure those programs will run seamlessly on the validation server, so long as synthetic data record layouts are identical to the confidential data, and receive statistically valid results. For instance, until September 2022, the US Census Bureau used to support validation servers for two experimental synthetic databases via the Synthetic Data Server at Cornell University: the Synthetic Longitudinal Business Database (Kinney, Reiter, and Miranda 2014)[6] and the Survey of Income and Program Participation's (SIPP) Synthetic Beta Data Product (Benedetto, Stanley, and Totty 2018).[7]

The idea of tiered access has been proposed in several public policy discussions. For instance, the Foundations for Evidence-Based Policymaking Act of 2018 (often called the "Evidence Act") "requires [federal] agency data to be accessible and requires agencies to plan to develop statistical evidence to support policymaking" (see https://www.congress.gov/bill/115th-congress/house-bill/4174). The Evidence Act also calls for the establishment of a National Secure Data Service, which could serve as a host for data, validation servers, and verification servers. In 2022, the Advisory Committee on Data for Evidence Building, a congressional committee charged with "reviewing, analyzing, and making recommendations on how to promote the use of Federal data for evidence building," released a final

[6]See the "Synthetic Longitudinal Business Database (SynLBD)," at https://www.census.gov/programs-surveys/ces/data/public-use-data/synthetic-longitudinal-business-database.html (accessed on January 15, 2024).

[7]See the "Synthetic SIPP Data," at https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html (accessed on January 15, 2024).

report recommending the creation of tiered access systems.[8] During the summer of 2023, America's Datahub Consortium, a collaboration sponsored by the National Center for Science and Engineering Statistics within the National Science Foundation, posted nine requests for solutions. Several of these requests involve creating demonstration products to support a tiered access model and explore the use of verification metrics in validating estimates produced from public data.[9]

Implementing Formal Privacy for Tiered Data Access

Although the Synthetic Data Server at Cornell University provided access to confidential data, it was not automated. The demand for the service often exceeded available staff time, causing long delays for approval. Another drawback is that the manual review processes involve high-level staff reviewing program code and output to identify potential disclosures of confidential data. This manual review process is time-consuming, based on human subjectivity, and imperfect. Also, because the staff reviewed proposals one at a time, they did not appropriately account for the cumulative disclosure risk over time.

A well-designed and automated validation server system could provide consistent and robust privacy protection with little or no human review, which is both safer and less labor-intensive than past manual review of research programs that involve subjective human review. To complement the synthetic SOI public use file, the Urban Institute and Statistics of Income Division at the IRS are developing an automated validation server that uses differentially/formally private methods to release statistically valid results with privacy protections (Barrentios et al. 2023; Taylor et al. 2021). Before delving into the practical challenges of such a system, I will provide a general background on differential/formal privacy.

Making a Noisy Case for New Privacy Definitions

The illustrative socioeconomic dataset provided above showed how implementing statistical data privacy methods to release public data and statistics is a viable privacy-protecting alternative for data users who are unable to access confidential data directly. Yet, like direct data access, these traditional approaches have their limitations. One challenge is predicting the behavior of data intruders, making it difficult to determine what information should be considered sensitive. A notorious example of how a malicious actor could gain private information from seemingly anonymous data is the Netflix Prize. Based on the movies that Netflix users liked in the past, Netflix wanted to recommend future movies that people will

[8] See the "Advisory Committee on Data for Evidence Building (ACDEB) releases final report (10.14.22)," at https://www.aeaweb.org/forum/3174/advisory-committee-evidence-building-acdeb-releases-report (accessed on July 11, 2023).

[9] See "Opportunities and Pending Awards," at https://www.americasdatahub.org/opportunities/ (accessed on July 11, 2023).

rate highly. Thus, back in 2006, Netflix organized a competition with a $1 million prize for an algorithm that would look at past movie ratings and then lead to users ranking the recommended movies more highly—specifically, developing a recommendation system to beat the existing Netflix algorithm by 10 percent. For the contest, Netflix provided a dataset comprising of about 100 million movie ratings from nearly 500,000 anonymous users.

However, instead of using the dataset to improve movie rating predictions, Narayanan and Shmatikov (2008) showed that they could reidentify the supposedly anonymized records in the dataset. They achieved this by using publicly available information from ratings on the Internet Movie Database (IMDb), an Amazon.com owned online database of actors, movies, TV series, and video games. Although knowing a person's movie rating may appear innocent, a data attacker could use preferences about movies to infer more sensitive aspects of people's lives, such as their sexual orientation or political preferences. Netflix did award the $1 million prize, but faced with lawsuits over privacy concerns, the company cancelled its planned follow-up contest in 2010 (as reported in Lohr 2010).

In a more recent example, the *New York Times* acquired a large cell-phone tracking dataset that contains only time and location information. In the paper, Thompson and Warzel (2019) wrote an article about how they successfully identified an individual based on deviations from their usual work routine, specifically noting a shift from commuting to Microsoft to commuting to Amazon.

These examples illustrate how traditional methods of protecting privacy can fall short. Data curators and privacy experts often struggle to predict what resources a data intruder has at their disposal. Does a data intruder have access to other databases or considerable computing power? How could a data intruder use these resources to exploit a combination of public data and other publicly available statistics? Without predicting the data intruder's behavior, data curators and privacy experts will have difficulties accounting the cumulative disclosure risk from a series of data releases, occurring each time a public dataset or statistic is released based on confidential data, which might gradually lead to revealing personal information.

Another drawback of traditional statistical data privacy methods is their lack of transparency, which hinders reproducibility and replicability. The traditional methods of protecting individual privacy often rely on the concept of "security through obscurity," where parts of these methods are concealed to prevent clever data intruders from reverse engineering and recreating the confidential data. However, as various professional and research organizations advocate for reproducibility and replicability in research,[10] traditional statistical data privacy methods will fall short on this dimension too.

[10] Examples include the American Economic Association (in its "Data and Code Availability Policy," at https://www.aeaweb.org/journals/data/data-code-policy) and the American Statistical Association (in its "ASA Journal Policies on Data Sharing and Reproducibility," at https://www.amstat.org/publications/q-and-as/asa-journal-policies-on-data-sharing-and-reproducibility).

Introducing Formal Privacy

The relatively new mathematical concept of differential privacy addresses the challenges that traditional statistical data privacy methods face, revolutionizing the field of data privacy and confidentiality. A report done for the US Census Bureau defined these privacy definitions, or "formal privacy" definitions, as a subset of statistical data privacy methods that provide "formal and quantifiable guarantees on inference disclosure risk and known algorithmic mechanisms for releasing data that satisfy these guarantees" (JASON 2022, p. 41). In other words, what makes a privacy definition formally private is broadly the ability to (1) quantify and adjust the privacy-utility trade-off (typically through parameters); (2) rigorously and mathematically prove the maximum privacy loss that can result from the release of information; and (3) compute total disclosure risk or privacy loss from multiple individual information releases (Bowen and Garfinkel 2021). I will refer to differential privacy and related privacy definitions that satisfies these characteristics as "formal privacy."[11]

Although noise addition has existed before formally private methods (as in the example earlier in the paper), the ability to account for the total privacy loss is crucial in ensuring strong privacy protection. Formal privacy methods are akin to using a debit card with a predetermined budget, whereas traditional statistical data privacy methods are like a limitless credit card. In both cases, there is a cumulative cost of associated purchases, but only the debit card requires constant monitoring of the remaining balance. In both traditional and formal privacy settings, data curators must restrict the type and quantity of queries made to the data. However, a formally private framework requires data curators to exercise diligence in tracking the usage of data to ensure privacy protections.[12]

Since its inception, differential privacy has been the most well-known formal privacy definition (Dwork et al. 2006). As mentioned earlier, differential privacy adheres to strict mathematical conditions to be considered differentially private, which is not a statement or description regarding data itself (in other words, differentially private methods are data agnostic). Thus, differential privacy does not make assumptions about what a data intruder considers sensitive information, nor what external data or computational power the intruder has access to, now or in the future. Differential privacy instead assumes the worst-case scenario that the data intruder has information on every record in the confidential data except one, unlimited computational power, and the record that the intruder has no information on is the most extreme possible record (or an extreme outlier) that could alter the target statistic or information that a data curator wants to release publicly.

This worst-case scenario assumption also enables data curators to disclose details of differentially private and formally private methods. For instance, privacy

[11] A factor that led to the generalized term of formal privacy is that privacy experts developed alternative versions of differential privacy that "relaxed" differential privacy's strong privacy guarantees (that is, the worst-case scenario conditions listed earlier). Although differential privacy is a more common definition, most practical applications use a relaxation or an alternative formally private definition, such as the 2020 Census (Bowen, Williams, and Pickens 2022).

[12] This analogy is borrowed from Bowen and Garfinkel (2021).

researchers can publish the privacy parameter values often referred to as privacy-loss budgets. This openness about methods contrasts with traditional statistical data privacy methods, such as when the US Census Bureau implemented "data swapping," a statistical data privacy method that involves exchanging observations with similar variable characteristics, from the 1990 Census to the 2010 Census. Throughout that time, the Census Bureau did not report the swapping rate, due to the risk of a data intruder reverse-engineering the method.

Purchasing Statistics with a Privacy-Loss Budget

The other key feature of formal privacy definitions is their ability to account for the cumulative disclosure risks or privacy-loss with the public release of information derived from confidential data. These definitions use the concept of a privacy-loss budget that is typically represented as ϵ. (There are additional privacy parameters for various definitions, such as δ and ρ. For simplicity, I will refer to the privacy budget as ϵ.) The data curator can treat the privacy-loss budget or privacy parameter like a knob to adjust the trade-off between privacy and utility when releasing a statistic of dataset publicly. This means the data curator must set the privacy-loss budget before publishing any information publicly and must track the budget or risk exhausting the budget prematurely. Similar to a financial budget, if the privacy-loss budget is exhausted, then no more information from the confidential data would be released.

Earlier in the article, I stated how a perfectly protected dataset is one that is never released, and a perfectly useful dataset is one that is released unaltered. Formal privacy frameworks can explain these extreme scenarios with ϵ. When ϵ becomes very large (approaching infinity), the released dataset is unaltered, but has no privacy. When ϵ becomes very small (approaching zero), the released dataset has maximum privacy, but no utility. Finding the balance between the two extremes is easier with the privacy-loss budget, because the data curator can increase the privacy-loss budget if they desire a more accurate statistic (but less privacy) or want more privacy (but less accuracy).

Data curators can also allocate the privacy-loss budget over several public datasets and statistics. As an example, imagine the privacy-loss budget is a monthly budget for household expenses (like spending on housing, groceries, utilities, and transportation). Some people might want to allocate their budget equally to each category of monthly expense; others might want to allocate more of their monthly budget to groceries than transportation. Similarly, some data curators might prioritize releasing multiple statistics, while other data curators might allocate the full privacy budget to allow the release of a more detailed microdata. The privacy-loss budget empowers data curators on how they can allocate and account for each individual release of information, while maintaining the overall budget for the system.[13]

[13] Analogy is borrowed from Bowen, Williams, and Pickens (2022).

Calculating the Global Sensitivity of Statistics

Most formally private methods use the concept of "global sensitivity," which describes how resistant a statistic is to the presence of outliers, tying into the "spending power" of the privacy-loss budget. If a statistic is more robust or resistant to outliers, less privacy-loss budget is needed to release a more accurate statistic. For the converse, a statistic that is less robust or resistant to outliers will require more privacy-loss budget to release a more accurate statistic.

The concept of global sensitivity works by quantifying how much an output can change with the addition or removal of the most extreme possible record that could possibly exist in the population. This means that regardless of whether that record is present in the data, we must consider that the record could be in the data—this is why formally private methods are data agnostic. Simply put, formally private methods use global sensitivity to account for any possible version of the data that could exist, protecting both against future data releases and new technologies.

Imagine the data that need protection contains socioeconomic information, and the question being asked is, "What is median wealth?" Within a formally private framework, it must consider the most extreme possible record that could exist in any given data that have demographic and financial information. In this example, that person is Elon Musk, who was the wealthiest person in the world at the end of 2023 according to *Forbes* magazine.[14] If Musk is present or absent in the data, the median wealth should not change much. This means a more accurate answer could be provided with fewer alterations to the median income statistic, because the statistic is less sensitive to the extreme outlier, Musk. In contrast, consider the question: "What is the average wealth?" Unlike the previous statistic, the answer to that statistic would drastically change if Musk were present or absent from the data. To protect the extreme case at a given level of privacy-loss budget, a formally private algorithm would need to provide a significantly less accurate answer by altering the statistic more.[15]

Highlighting Practical Challenges of an Automated Validation Server with Formal Privacy

With the basics of formal privacy in place, we can imagine how an automated validation server could enable more data users to have access to the underlying administrative data, without placing additional burdens on government staff. At a high level, an automated validation server using formal privacy would require the data curator to monitor a privacy budget, allowing the curator to track the cumulative effect of several submitted analyses from data users.

While the concept of formal privacy holds promise for automated validation and verification servers, it poses several implementation challenges. To help explain these challenges, I will discuss an example: a collaboration between the

[14] For details, see "The World's Real Time Billionaires," available at https://www.forbes.com/real-time-billionaires (accessed on January 15, 2024).

[15] Analogy is borrowed from Bowen, Williams, and Pickens (2022).

Urban Institute and Statistics of Income Division at the IRS to create an automated validation server applying formally private methods that enables safe access to confidential administrative tax data.

One challenge has been the discovery that the formal privacy field is not as far along as it may have seemed a few years ago. Barrientos et al. (2023) conducted a feasibility study on the state-of-the-art formally private methods to release summary statistics and regression coefficients, with evaluations on administrative tax data from the Statistics of Income Division at the IRS and survey data from the Census Bureau. They found that formally private methods for summary statistics perform well for small privacy-loss budgets, which means that for this purpose, formal privacy provides more "bang for your buck." In contrast, the formally private regression methods require much larger privacy-loss budgets. Additionally, very few of these formally private methods provided standard errors for regression coefficients, which are essential for most social science research. The authors highlighted that more research is needed to develop methods that are robust to data types that do not involve normal distributions, and to calculate the uncertainty around the estimates. There is a lot of research on how to implement formally private methods for prediction or the outcomes of regression models.

Based on research studies and discussions with privacy experts, data users, and the Urban Institute project team, Snoke et al. (2024) identified five types of incompatibilities between current practices in statistical data analysis and data privacy—specifically, estimates for traditional statistical inference, control or nuisance variables, assumptions on the range of the data or other assumptions, performing exploratory data analysis, and limited queries and the privacy budget. I will discuss the last two.

Most data users need to perform exploratory data analysis and subsequent analyses. In many cases, data users may be unsure of the specific analyses they want to conduct until they have access to the data. Because data users will have a limited privacy budget, a validation server can provide a disincentive for undesirable research practices like p-hacking (that is, searching through multiple, alternative specifications to find one that meets the criterion for statistical significance). However, concerns remain regarding how a formally private validation server can handle other research practices, such as multiple testing or when a journal reviewer requests that alternative models be applied.

This challenge sheds light on a broader issue for formal privacy: setting privacy budgets for data users. How should data curators allocate privacy budgets to different data users? How can data users determine their model specifications without depleting their entire budget? How do data users conduct robustness checks without exhausting their allocated resources? A related issue arises when multiple data users submit the same analysis. Suppose data user A queries a statistic, and two months later, data user B queries the same statistic on the same part of the data. How should the validation server handle this scenario? There are two general options: (1) treat queries A and B as separate queries (with two different results) and charge both users their respective privacy-loss budgets (which could be the

same or different amounts); or (2) use the privacy-preserving result from A for B. Both have pros and cons.

The former approach avoids the conflict of notifying data user B that data user A has already conducted the analyses. After all, researchers may not want others to know what kind of analyses they are conducting for various reasons, such as to avoid being scooped. But submitting the queries separately would result in data user A and data user B having slightly different answers to the same statistic. This creates communication and education problems in explaining to both data users that their answers are valid (a problem to be discussed further in the next section).

The latter approach would avoid the confusion of having two different answers. The data users could also split the privacy-loss, reducing the cost to their respective privacy-loss budgets. A drawback is that both data users would be informed that another has conducted a similar analysis. An additional complication is if data user B wants a more accurate result and, therefore, wants to spend a higher privacy-loss budget than data user A, who does not want to exceed a certain privacy-loss amount.

Measuring the Privacy-Loss Budget

Although the concept of privacy-loss budgets addresses the ad hoc nature of traditional statistical privacy methods, it brings other complicated issues. The incompatibility problems discussed earlier are just some of many others needing to be addressed, such as most statistical data privacy methods require the data user to make assumptions without access to the data (Snoke et al. 2024).

The challenge of bringing formal privacy directly to real-world datasets is an ongoing subject of applied research. As an example of a central problem, one of my colleagues working at the Urban Institute asked, "Does it make sense to think of a unit of privacy-loss budget as having the same value across statistics, functions, and runs?"

One might assume that a given unit of privacy-loss should be considered equally valuable across different statistical analyses, but, in reality, this seems unlikely. As a metaphor, suppose you have $20. Regardless of where you are in the United States, that $20 holds the same face value. However, in one area, it might be sufficient to buy a decent meal at a specific restaurant, while in another area, it may not be enough. Furthermore, what defines a good meal may differ across people. Similarly, a certain quantity of privacy may seem satisfactory in some contexts, but not in others, and it may satisfy some people, but not others.

When I shared this analogy with another colleague, the response was, "I like the metaphor that you don't always get what you pay for. But, to me, the currency is more like Martian bucks instead of US dollars." My colleague's point highlights the fact that the field of data privacy and confidentiality has no inherent understanding of the actual value of the privacy-loss budget currency. All we know is that increasing the privacy-loss budget should result in more accurate analysis or produce a higher quality dataset. Because we do not know what a privacy-loss budget truly affords us, it might as well be from another world.

A New Landscape of Data Privacy

Data curators, stakeholders, privacy experts, and data users all face challenges concerning a desire to expand the use of government data, especially in building linkages between administrative and survey data. Unless and until data privacy is securely protected, such data are likely to be available only narrowly, or not at all.

Educating the Data User Community

Little is known about the expectations and needs of data users in general—let alone their understanding and perceptions of formal privacy. Williams et al. (2023) conducted a convenience sample survey of economists from the American Economic Association on their baseline knowledge about differential/formal privacy, attitudes toward differentially/formally private frameworks, types of statistical methods that are most useful to economists, and how the injection of noise under formal privacy would affect the value of the queries to the user. At a high level, the survey found that most economists are unfamiliar with formal privacy and differential privacy (and if they know about it, they are skeptical). Instead, economists rely on simple methods for analyzing cross-sectional administrative data but have a growing need to conduct more sophisticated research designs, and economists have low tolerance for errors, which is incompatible with existing formal privacy definitions and methods.

The results from the Williams et al. (2023) survey are not surprising. In general, traditional statistical data privacy methods are more intuitive and easier to explain, such as why data curators should remove unique records. In contrast, formally private methods are more complex and lack an intuitive definition.

Although there has been an explosion of new communication materials to explain formal privacy and other data privacy concepts,[16] such efforts are trying to fill a chasm and we are not even close. To put it into perspective, if we asked random economists to recommend their favorite education or communication materials about, say, machine learning or artificial intelligence, many would have a favorite book or blog series in mind. They may even have suggested materials that are more focused on concepts, or a perspective from a certain field, or on coding. If we asked random economists the same question, but for data privacy in the context of safe access to administrative and survey data, they likely would have few recommendations or even none at all.

One way to address the lack of education and communication materials is to teach the next generation and increase the number of those in the field. Yet despite the need for data privacy education, most higher education institutions do not offer dedicated courses on the topic. When data privacy is taught, it is typically at the graduate level within computer science departments. Some undergraduate professors who research data privacy and confidentiality may introduce these topics in

[16] One of my favorites is a video created by minutephysics for the US Census Bureau, available at "Protecting Privacy with MATH (Collab with the Census)," https://www.youtube.com/watch?v=pT19VwBAqKA (accessed on June 21, 2023).

seminar courses, but they are not usually stand-alone courses. As a result, individuals with technical backgrounds outside of computer science, such as economists, are greatly underrepresented in this important area of study. Therefore, other departments outside of computer science should consider hosting their own statistical data privacy courses or incorporating these concepts into existing courses. When integrating these statistical data privacy concepts, professors can encourage students to consider the legal, social, and ethical implications of data privacy, ethics, and equity. They can also delve into the principles of data guardianship, custodianship, and data permissions (Williams and Bowen 2023).

Addressing Data Equity in Data Privacy

The methods used to protect individuals' information do not always have an equal impact on all groups represented in the data. A published dataset might ensure the privacy of people who are the majority in the dataset but fail to ensure the privacy of those in smaller groups. Similarly, alterations to the data may be more useful for learning about some groups more than others. Ultimately, how entities collect and share data can have varying effects on underrepresented groups of people.

Although there are many discussions on data equity *and* data privacy, few conversations focus on equity *in* privacy. In light of this, Bowen and Snoke (2023) developed a guide as part of the "Do No Harm Guides" series. This fourth installment of the series focuses on exploring the current state of equity-focused work in statistical data privacy. The authors conducted interviews with nine experts in privacy-preserving methods and data-sharing, including researchers and practitioners from academia, government, and industry sectors with diverse technical backgrounds. The authors asked about the experience of these experts in implementing statistical data privacy methods and how they define equity in the context of privacy, among other topics. The authors then created an illustrative example to highlight potential disparities that can result from applying various statistical data privacy concepts (including suppression, synthetic data, and differential privacy) without an equitable workflow. Here are some of their key takeaways: do not treat equity as a separate field of study; work with groups represented in your data; and there is no methodological silver bullet.

Engaging with Data Privacy Issues

There are a few prominent options for learning about data privacy methods and becoming involved in these topics besides becoming a privacy researcher. For instance, the Joint Program in Survey Methodology at the University of Maryland has been offering a course on synthetic data.[17] The Urban Institute offered an all-day course at the 2023 Joint Statistical Meetings, where the instructors introduced the

[17] See "Synthetic Data: Balancing Confidentiality and Quality in Public Use Files," a course by Joerg Drechsler and Jerome P. Reiter, at https://jpsm.umd.edu/academics/jpsm-short-course-2023-24-schedule (accessed on June 21, 2023).

basics of data privacy.[18] The Urban Institute has also offered similar trainings for the Bureau of Economic Analysis, Allegheny County, and the Statistics of Income Division.

There is currently no dedicated conference focused on the intersection of data privacy and public policy, but interest in the field is growing. In 2023, the National Bureau of Economic Research[19] and National Institute of Statistical Sciences[20] hosted separate data privacy workshops that brought together privacy experts and data users. Attendees from these workshops are currently organizing the first ever Privacy and Public Policy Conference in 2024 with the goal "to foster and enhance collaboration among privacy experts, researchers, data stewards, data practitioners, and public policymakers."[21] With the recent surge of venues, the time is obviously ripe to help shape the future of data privacy, make meaningful contributions to its policy debates, and ensure the responsible representation of people in data.

[18] See "Introduction to Data Privacy and Data Synthesis Techniques," acourse by Aaron R. Williams and Claire McKay Bowen, at https://ww2.aievolution.com/JSMAnnual/index.cfm?do=ev.pubSearchEvents (accessed on June 21, 2023).

[19] See "Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches, and their Consequences, Spring 2023," hosted by the National Bureau of Economic Research, at https://www.nber.org/conferences/data-privacy-protection-and-conduct-applied-research-methods-approaches-and-their-consequences (accessed on June 21, 2023).

[20] See "IOF Workshop: Advancing Demographic Equity with Privacy Preserving Methodologies," hosted by the National Institute of Statistical Sciences, at https://www.niss.org/events/iof-workshop-advancing-demographic-equity-privacy-preserving-methodologies (accessed on June 21, 2023).

[21] See "Privacy and Public Policy Conference," at https://privacypublicpolicy-conference.github.io/website/ (accessed on January 15, 2024).

References

Barrientos, Andrés F., Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. 2018. "Providing Access to Confidential Research Data through Synthesis and Verification: An Application to Data on Employees of the US Federal Government." *Annals of Applied Statistics* 12 (2): 1124–56.

Barrientos, Andrés F., Aaron R. Williams, Joshua Snoke, and Claire McKay Bowen. 2023. "A Feasibility Study of Differentially Private Summary Statistics and Regression Analyses with Evaluations on Administrative and Survey Data." *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2023.2270795.

Belkin, Douglas. 2020. "No Place to Hide: Colleges Track Students, Everywhere." *Wall Street Journal*, March 5. https://www.wsj.com/articles/the-many-ways-college-students-may-be-tracked-on-campus-11583354852.

Benedetto, Gary, Jordan C. Stanley, and Evan Totty. 2018. "The Creation and Use of the SIPP Synthetic Beta v7.0." US Census Bureau Working Paper.

Bierbrauer, Felix J., Pierre C. Boyer, and Andreas Peichl. 2021. "Politically Feasible Reforms of Nonlinear Tax Systems." *American Economic Review* 111 (1): 153–91.

Bowen, Claire McKay, Victoria Bryant, Leonard Burman, John Czajka, Surachai Khitatrakun, Graham MacDonald, Robert McClelland et al. 2022a. "Synthetic Individual Income Tax Data: Methodology, Utility, and Privacy Implications." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Maryline Laurent, 191–204. Cham: Springer Nature.

Bowen, Claire McKay, Victoria Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland,Livia Mucciolo, Madeline Pickens, and Aaron R. Williams. 2022b. "Synthetic Individual Income Tax Data: Promises and Challenges." *National Tax Journal* 75 (4): 767–90.

Bowen, Claire McKay, and Simson Garfinkel. 2021. "Philosophy of Differential Privacy." *Notices of the American Mathematical Society* 68 (10): 1727–39.

Bowen, Claire McKay., and Joshua Snoke. 2023. *Do No Harm Guide: Applying Equity Awareness in Data Privacy Methods*. Washington, DC: Urban Institute.

Bowen, Claire McKay, Aaron R. Williams, and Madeline Pickens. 2022. "Decennial Disclosure." Washington, DC: Urban Institute.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79.

DeBacker, Jason, Richard W. Evans, and Kerk L. Phillips. 2019. "Integrating Microsimulation Models of Tax Policy into a DGE Macroeconomic Model." *Public Finance Review* 47 (2): 207–75.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography*, edited by Shai Halevi and Tal Rabin, 265–84. New York: Springer Berlin Heidelberg.

Hu, Jingchen, and Claire McKay Bowen. 2023. "Advancing Microdata Privacy Protection: A Review of Synthetic Data Methods." *Wiley Interdisciplinary Reviews: Computational Statistics* 16 (1): e1636.

JASON. 2022. "Consistency of Data Products and Formally Private Methods for the 2020 Census." US Census Bureau. https://irp.fas.org/agency/dod/jason/census-privacy.pdf (accessed on June 21, 2023).

Kinney, Satkartar K., Jerome P. Reiter, and Javier Miranda. 2014. "Synlbd 2.0: Improving the Synthetic Longitudinal Business Database." *Statistical Journal of the IAOS* 30 (2): 129–35.

Lohr, Steve. 2010. "Netflix Cancels Contest after Concerns Are Raised about Privacy." *New York Times*, March 13. https://www.nytimes.com/2010/03/13/technology/13netflix.html.

Matthews, Gregory J., and Ofer Harel. 2011. "Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy." *Statistics Surveys* 5: 1–29.

McClelland, Robert, Daniel Berger, Alyssa Harris, Chenxi Lu, and Kyle Ueyama. 2019. *The TCJA: What Might Have Been*. Washington, DC: Urban-Brookings Tax Policy Center.

McKenna, Laura, and Matthew Haubach. 2019. "Legacy Techniques and Current Research in Disclosure Avoidance at the US Census Bureau." US Census Bureau Working Paper CED-WP-2019-005.

Nagaraj, Abhishek, and Matteo Tranchero. 2023. "How Does Data Access Shape Science? Evidence from the Impact of US Census's Research Data Centers on Economics Research." NBER Working Paper 31372.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets." In *2008 IEEE Symposium on Security and Privacy*, 111–25. Piscataway, NJ: IEEE.

Rubin, Donald B. 1987. "Multiple Imputation for Nonresponse in Surveys." New York: Wiley.

Snoke, Joshua, Claire McKay Bowen, Aaron R. Williams, and Andrés F. Barrientos. 2024. "Incompatibilities between Current Practices in Statistical Data Analysis and Differential Privacy." *Journal of Privacy and Confidentiality*. https://arxiv.org/pdf/2309.16703.pdf.

Taylor, Silke, Graham MacDonald, Kyle Ueyama, and Claire McKay Bowen. 2021. *A Privacy-Preserving Validation Server Prototype*. Washington, DC: Urban Institute.

Thompson, Stuart A., and Charlie Warzel. 2019. "Twelve Million Phones, One Dataset, Zero Privacy." *New York Times*, December 19. https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html.

Williams, Aaron R., and Claire McKay Bowen. 2023. "The Promise and Limitations of Formal Privacy." *Wiley Interdisciplinary Reviews: Computational Statistics* 15 (6): e1615.

Williams, Aaron R., Joshua Snoke, Claire McKay Bowen, and Andrés F. Barrientos. 2023. "Disclosing Economists' Privacy Perspectives: A Survey of American Economic Association Members on Differential Privacy and Data Fitness for Use Standards." Paper presented at the Data Privacy Protection and the Conduct of Applied Research: Methods Approaches and Their Consequences NBER conference, Cambridge, MA, May 4.