BROOKINGS

APRIL 2026

A FRAMEWORK FOR EVALUATING INNOVATION IN FEDERAL ECONOMIC STATISTICS

Claire McKay Bowen

This report is part of The Economic Indicators Initiative, a think tank collaborative dedicated to producing consensus-building research on improving key U.S. economic indicators. Learn more online at https://www.brookings.edu/economic-indicators-initiative/

AUTHOR NOTES AND ACKNOWLEDGEMENTS

Claire McKay Bowen is a senior fellow in the Tax and Income Supports Division and leads the Data Governance and Privacy Team at the Urban Institute.

DISCLOSURES

The Brookings Institution is committed to quality, independence, and impact. We are supported by a diverse array of funders. In line with our values and policies, each Brookings publication represents the sole views of its author(s).

Introduction

Federal economic statistics are essential to evidence-based policymaking in the United States, informing decisions on issues such as job losses or sudden shifts in demand for goods. Yet their production faces significant challenges, including declining survey response rates, constrained budgets, and growing expectations for timely, granular data. This paper examines how these data are produced across the full data life cycle, which consists of six phases: (1) collection and acquisition, (2) storage, (3) sharing and transfer, (4) analysis, (5) dissemination, and (6) destruction or archival.

Understanding these phases is critical because challenges—and opportunities for innovation—arise at every stage. To frame this discussion, I define the core values of data governance as ensuring that data are accurate, accessible, private, and usable. Innovation should then introduce new ideas or changes that strengthen the federal statistical system's ability to produce economic data that meet these core values for supporting better evidence-based decision-making, ultimately improving the well-being of our communities and nation.

- **Accuracy** ensures that society has high-quality data and statistics that meaningfully represent the data subjects. Inaccurate data can lead to flawed analyses, poor decisions, and misrepresentation.
- **Accessibility** addresses what data and statistics are available, to whom, and under what conditions. Data should be accessible to those who need it for legitimate purposes.
- **Usability** means data must be understandable, actionable, and fit for purpose. Even high-quality data can be ineffective if users cannot interpret or apply it.
- **Privacy** safeguards sensitive information from illegitimate access or misuse. Numerous laws and regulations—from local to federal—require data curators to protect privacy rights and maintain public trust.

These values are not discussed in detail later in the paper but serve as a lens for evaluating the persistent challenges and innovations highlighted in the following sections. Specifically, the paper identifies fifteen persistent challenges in producing high-quality economic data across the six life cycle phases. These include integrating administrative and private sector datasets, ensuring representation of hard-to-reach populations, balancing privacy protection with data utility, and modernizing metadata. To illustrate potential solutions, the paper highlights four recent innovations:

1. Re-Engineering Statistics using Economic Transactions (RESET), which applies transaction-level data to improve GDP and price measurement;
2. National Experimental Well-being Statistics (NEWS), which blends survey and administrative data to refine income and poverty estimates;
3. Comprehensive Income Dataset (CID), which links multiple sources to fill gaps in measures of economic well-being; and
4. Safe Data Technologies, which applies privacy-enhancing technologies for secure access to confidential administrative tax data.

Together, these examples show how technical and policy innovations can strengthen the federal statistical system's ability to deliver cost-efficient, high-quality economic statistics that meet evolving societal needs. Not all proposed innovations will meet the core data governance values. The recent innovations must also address gaps in workforce skills, infrastructure, AI integration, and data archiving to uphold these core values. The paper concludes by calling for sustained modernization, cross disciplinary training, and adaptive policies to ensure federal economic statistics remain robust, trusted, and fit for purpose in a rapidly changing technological and economic landscape.

Data life cycle challenges for federal economic statistics

Almost everyone works with or interacts with federal data and statistics—often without realizing it (Bowen and Williams 2025a). Even for those who are aware of federal data's ubiquity, most think only about the analysis or dissemination of data, rarely considering the other parts of the data life cycle. Each stage of that life cycle presents particular challenges, which also represent areas where potential solutions or innovations can be applied.

- **Phase 1: Data collection and acquisition:** designing and obtaining data from various sources
- **Phase 2: Data storage:** housing the collected data in physical or digital repositories
- **Phase 3: Data sharing and transfer**: distributing data between entities, individuals, or systems
- **Phase 4: Data analysis**: exploring, processing, and interpreting data to provide insights and trends that can inform decision-making
- **Phase 5: Data dissemination:** creating written, digital, and verbal products to relevant stakeholders
- **Phase 6: Data destruction and archival**: disposing or archiving of data that is no longer needed

I will use the decennial census as an example, which is one of the most important data collection efforts in the United States, to describe this cycle. As mandated by the Constitution,[1] census data subjects include every "whole number of persons" living in the United States and its five territories. Collecting data (phase 1) on such a massive scale requires numerous and varied data collectors—from methodologists who design the census to field agents who go door-to-door verifying responses. Next is storage and access (phase 2), which depends on the type of data requested. For confidential, micro-level census data, these datasets are only publicly released 72 years later after they are collected, as required by U.S. Code Title 44.[2] However, they may be made available to individuals sworn to protect confidentiality (i.e., those with Special Sworn Status) and who access the data in secure environments such as a Federal Statistical Research Data Center. [3] Depending on the data type, some datasets are transferred and shared (phase 3) with other entities to create different data products or stored in other secure facilities like ICPSR.[4] The analysis and dissemination stage (phases 4 and 5) is the responsibility of data users and practitioners. For example, census data are used by the Department of Education to determine Title I funding for schools and school districts with high percentages of students from low-income families (Moulton and Luong, 2023; National Academies of Sciences, Engineering, and Medicine, 2015). Finally, archival and destruction (phase 6) occur when data are no longer actively used. Decennial census data are archived and made publicly accessible after 72 years, as seen most recently with the release of the 1950 Census.[5]

At each stage of the data life cycle, there are several challenges that are often technical, legal, societal, and ethical in nature. Innovation must tackle these issues in ways that strengthen the four values of data governance—accuracy, accessibility, privacy, and usability—and ensure that everyone is responsibly represented. I outline the various challenges at each stage of the life cycle.

PHASE 1: DATA COLLECTION AND ACQUISITION

Before data collection even begins, any individual or entity interested in collecting data must first ask, "Why are we collecting these data, and who will use them?" These questions—and others—must be answered because the decision to collect or not collect and how data should be collected will shape how the data can be used downstream.

Challenge 1: Can leveraging administrative data, blending multiple data sources, or substituting government data with private-sector data mitigate the declining survey response rates?

Declining survey response rates raise critical questions about how official economic statistics can adapt. Commonly proposed strategies are leveraging administrative data, integrating multiple sources, or substituting government data with private-sector data. Each approach offers benefits and trade-offs. Surveys let agencies define data

fields and capture details beyond administrative records, but they are costly and face rising nonresponse rates. Administrative data are cheaper and readily available but are often limited to respondents in specific programs or activities. Government data are generally stable and may have public versions, while private data can be withdrawn at any time and often requires payment. However, private sources may include valuable measures that are difficult or expensive for the government to collect. Blending data sources like surveys and administrative records can provide a fuller picture of our communities and nation but poses challenges in integrating the different sources when not designed to be combined and can heighten privacy concerns.

Challenge 2: Are these data sources properly representing data subjects?

Even if a federal agency could collect the economic data through administrative processes, surveys, or third-party sources, ensuring proper representation of those who are hard to reach—such as those living in rural areas or those with low internet access in the United States—remains a huge challenge. Yet capturing these populations is critical, as these groups are often the focus of public policy and advocacy efforts, such as supporting our nation's farmers.

Challenge 3: How can we address the benefits and risks of data collection?

One reason for not collecting certain data is the risk that it could be misused, particularly when individuals or specific groups are easily identifiable. For example, after Pearl Harbor, the War Department—now the Department of Defense—requested 1940 Census data on Japanese Americans at the census tract and block levels to facilitate their placement in internment camps. This misuse and violation of data privacy have led some people today to refuse participation in the decennial census or other government-supported data collections or to advocate for removing certain questions, such as those about citizenship status (Lo Wang, 2018).

A similar situation is unfolding again in 2025, with Immigration and Customs Enforcement requesting taxpayer data to track undocumented immigrants (Nguyen, 2025). These examples stress the privacy risks, concerns, and challenges inherent in collecting sensitive data, and they bring us back to the fundamental questions posed at the start of this section: "Why are we collecting these data, and who will use them?" In other words, what harms might arise from collecting—or not collecting—specific data?

PHASE 2: DATA STORAGE

Because the majority of those who work with data only see the final product, I imagine most of us take the accessibility of information for granted. Yet the data that fuels everything from advanced research to everyday decision-making is, by many measures, underappreciated and insufficiently rewarded (Thessen et al., 2019). This situation—combined with how those in the data ecosystem value data infrastructure—creates incentives that fail to support proper storage and access, leading to persistent challenges.

Challenge 4: How can we expedite data cleaning and maintenance?

After collection, raw data often requires extensive cleaning to address inconsistencies, missing values, and other issues—a process that typically consumes most of the time in the data life cycle. Cleaning is critical for transforming raw data into a usable form, so decisions made during this stage can significantly affect downstream analysis. For instance, missing data may be imputed to create complete datasets for statistical purposes, though gaps can stem from collection errors or legal and societal constraints.

The Internal Revenue Service (IRS) Master File is a great example that demonstrates the complexity and challenges of data cleaning. This file contains over 140 to 150 million unedited tax returns, depending on the year (Bowen et al., 2022b). Although comprehensive, the file's size (i.e., millions of returns with thousands of variables) and inconsistencies (e.g., incorrect filings or late submissions) limit its usefulness for policy analysis. Instead, the Statistics of Income (SOI) Division at the IRS—one of the 13 principle federal statistical agencies—produces a smaller, curated dataset called the INSOLE (Individual and Sole Proprietor) of over 300,000 records (Bowen et

al., 2022a). This file is more manageable but only a fraction of the available information, leaving the vast potential for evidence-based research untapped.

Challenge 5: How to ensure data are FAIR?

Even well-cleaned data are useless if the information cannot be easily found. Metadata is essential for discoverability, reusability, and interoperability, yet most datasets—especially federal ones—often lack consistent metadata.[6] This gap across public and confidential datasets from local, state, territorial, tribal, and federal agencies was one reason the Open, Public, Electronic, and Necessary (OPEN) Government Data Act was included in the 2018 Foundations for Evidence-Based Policymaking Act, a bipartisan U.S. law aimed at modernizing federal data infrastructure, including data management and statistical efficiency. [7]

Ensuring that data include metadata that are findable, accessible, interoperable, and reusable (FAIR)[8] is also critical for preparing for increased use of AI. While several states have passed laws and a recent executive order established a national policy framework for artificial intelligence,[9] true readiness for the AI revolution requires that the underlying data and infrastructure are AI-ready. This means having proper metadata so that economic data are interpretable by both humans and machines and can be used as intended—rather than producing flawed or misleading results due to poor-quality inputs.

PHASE 3: DATA SHARING AND TRANSFER

Most of us who work with data were often taught (likely unintentionally) that data are everywhere and easily accessible, but that's not the reality. Typically, government data and statistics access comes in two forms: secure data access and public data access.

Secure access involves direct or restricted entry to sensitive data, often requiring government-issued laptops or travel to physical facilities such as Federal Statistical Research Data Centers.[10] These arrangements demand background checks and legal agreements—data use agreements (DUAs), nondisclosure agreements (NDAs), memorandums of understanding (MOUs)—that serve as guardrails against misuse. While this method is highly secure, it is also highly inaccessible. Physical travel can mean hundreds of miles for an individual; clearance requirements may exclude individuals without U.S. citizenship; and approval times can stretch into months or years. These delays are particularly problematic during national or global emergencies, such as the COVID-19 pandemic, when rapid data access on employment and labor was critical.[11]

Public data access, on the other hand, is what most people are familiar with: downloadable files or published statistics on government websites. Examples include population tables or unemployment rates. This approach is far more accessible—provided there is internet access—but accessibility comes with trade-offs. Federal agencies must apply strong privacy protections before release, like aggregating or coarsening values or excluding sensitive variables like high-income tax data to prevent identifiability. These measures safeguard privacy but can limit the usefulness of data for certain analyses.

Although this framework has existed for decades, it raises important questions about how data curators within federal agencies can safely share and transfer economic data, particularly across other federal agencies, other levels of government, or external partners.

Challenge 6: What guardrails are needed for responsible data access and sharing?

Secure facilities and legal agreements provide strong protections, but they also create technical, policy, and financial barriers. How do we strike the right balance between security and accessibility? What technical, legal, and procedural safeguards are essential to prevent misuse while enabling timely access?

Challenge 7: How does one handle consent?

Individuals may agree to share their data with one entity but not another. How do we honor those preferences when data moves across different jurisdictions or to external partners, such as states sharing education-to-workforce data to track labor mobility? What mechanisms can ensure that consent remains meaningful and enforceable throughout the data life cycle?

Challenge 8: Who decides whether data should be shared or transferred?

Data curators and privacy experts often collaborate to apply statistical data privacy methods or privacy-enhancing technologies—altering data just enough to protect identities while preserving the usability of data. But who ultimately makes the decision? Should these decisions rest with government agencies, independent boards, or community stakeholders? And how do we ensure transparency and accountability in this process?

PHASE 4: DATA ANALYSIS

There are many ways to use and analyze the information from federal data and statistics, and those choices can dramatically shape how results can be interpreted.

Consider these questions:

- What is the average student loan debt?
- At what year after being founded do most startup companies fail?
- Does the unemployment rate accurately reflect joblessness?
- What percentage of welfare recipients are able-bodied adults who don't work?

If I asked these questions to different data users, there is a non-zero probability that I will receive different answers. Each user may rely on different datasets, methods, and assumptions, which lead to varying results.

For instance, consider the first question: "What is the average student loan debt?" While it may seem straightforward, the answer can vary depending on the data source and when it was last updated. One article reports the average federal and private student loan debts as $39,375 and $42,673, respectively, based on data current as of September 2025.[12] Another source, last updated on February 2, 2026, lists those figures as $39,547 and $43,333.[13] You might assume I was intentionally searching for different numbers, but I received the first set of figures from Copilot and the second set from ChatGPT using the same simple prompt: "Answer the following question and provide the source for answering it: What is the average student loan debt?"

Context is key. When are the data updated? What are the sources? What is being analyzed? What assumptions are baked into the process? Decisions made during earlier stages of the data life cycle—collection, storage, transfer—shape analysis, even though analysis often consumes most of the spotlight. How do we acknowledge those upstream choices to ensure proper interpretation?

Challenge 9: How do we ensure standardized workflows?

Without consistent standards, analyses can diverge wildly. How do we create workflows that promote comparability and reliability across studies while still allowing flexibility for innovation?

Challenge 10: How do we promote transparency and responsible data use?

Do data users have the right data and metadata to make informed decisions about how to process and interpret data? Are there clear disclaimers, guidelines, or warnings to ensure outputs are used responsibly? Even small changes in data access—whether expanding or restricting it—can drastically alter the analysis results.

PHASE 5: DATA DISSEMINATION

Dissemination is not simply about releasing results; it is about communicating insights in ways that are accessible, accurate, and useful, while minimizing disclosure risks, potential harm, and misinformation. Achieving this balance requires careful attention to both technical and policy considerations.

Different audiences require different levels of depth—from technical papers for economists to social media posts for the general public. Defining the audience is essential. A "lay audience" might include community organizations, policymakers, tribal leaders, or advo-

cates, each with distinct needs for economic data and statistics. Without clarity, a data story intended for everyone risks becoming a story for no one.

Challenge 11: How to we make dissemination products accessible to the right audience?

One major challenge is designing products that meet these goals. How do we ensure that insights are conveyed clearly and responsibly? In the United States, this task is complicated by a patchwork of privacy laws—dozens at the federal level and additional regulations across states. These overlapping frameworks influence every stage of the process, from collection and storage to dissemination. When states attempt to share data across state lines, differing interpretations of these laws add further complexity, sometimes even among legal experts.

Challenge 12: How do we prevent multiple narratives from the same dissemination product?

Another challenge is anticipating and managing multiple narratives that can emerge from the same dataset. Data are rarely neutral. Policymakers, researchers, advocates, and the public may frame the same findings in very different ways. Understanding these dynamics is critical because dissemination decisions can shape public discourse and policy outcomes.

These considerations raise important questions for the field. How can dissemination strategies prioritize accessibility, accuracy, and privacy? How do we anticipate and manage competing narratives? What frameworks best balance public good with potential harm? And how do we tailor communication to diverse audiences without diluting the message? Addressing these questions is central to designing data products that truly serve their intended purpose.

PHASE 6: DATA ARCHIVAL AND TERMINATION

Within the data life cycle, a critical decision must be made: To archive or not to archive data? This decision is not merely technical—it raises ethical, practical, and transparency-related questions that we must navigate carefully.

Challenge 13: Should we archive or terminate the data?

When is it appropriate to archive data, and when should it be securely destroyed? While termination may seem like the default, there are cases where destroying data could be unethical. For example, eliminating datasets might disproportionately impact certain subpopulations, forcing them to participate repeatedly in future studies, creating survey fatigue or eroding trust (Chicago Beyond, 2019). Archiving can reduce respondent fatigue and preserve valuable insights, but it also introduces long-term storage and privacy considerations. How long should data be retained before destruction? What standards should guide these timelines?

Challenge 14: How do we ensure transparency without harming data collection?

Transparency is a cornerstone of ethical research, but it comes with trade-offs. Should participants be informed at the point of collection that their data will eventually be deleted? This aligns with principles like the Right to Be Forgotten from General Data Protection Regulation[14] and California Consumer Privacy Act.[15] However, overly emphasizing deletion policies could discourage participation. How do we strike a balance of providing clarity without creating fear or mistrust?

Challenge 15: Do we create time-dependent triggers and persistent record keeping?

Even when data are destroyed, should there be a record of its existence? Persistent metadata or summary statistics and other information can help maintain accountability and historical context without compromising privacy. What minimum information should be preserved to document that data once existed, and how should these records be managed over time?

These decisions are rarely static. Data landscapes evolve and so do community perspectives. Researchers must continually engage with stakeholders to weigh the trade-offs between privacy, security, transparency, and utility. Destroying data may protect individual rights, but such a practice can also create gaps that lead to repeated burdens on certain groups. Conversely, indefinite retention raises risks of misuse and breaches. The solution lies in ongoing conversations and adaptive policies.

Innovation for producing federal economic data and statistics

Despite these challenges, recent federal innovations grounded in the core data governance values— accuracy, accessibility, privacy, and usability—are reshaping how the data community and ecosystem may view the future economic data and statistics. Earlier, I defined "innovation" as new ideas or changes that strengthen the federal statistical system's ability to produce cost-efficient, high-quality economic data and statistical products that respond quickly and support evidence-based decisions, improving community and national well-being throughout the data life cycle. The following section highlights examples of recent innovations (within the past five years since the publication of this paper) that align with the data governance values, discussing their motivation, how they address one or more of the fifteen challenges, and any new or persistent challenges.

A full methodological review is beyond the scope of this paper; readers are encouraged to read the technical reports, peer-reviewed publications, and other dissemination products for detailed information.

RE-ENGINEERING STATISTICS USING ECONOMIC TRANSACTIONS

A collaborative project with the University of Maryland, University of Michigan, and the U.S. Census Bureau called Re-Engineering Statistics using Economic Transactions (RESET) "...aims to provide the architecture for re-engineering official economic statistics—literally to build key measurements such as GDP and consumer inflation from the ground up."[16] Beyond addressing the persistent challenge of declining survey response rates, a major motivation is to reconsider how key economic indicators—such as the National Income and Product Accounts (NIPAs)—should be measured considering the significant transformations in the economy in recent years. Today, everything from purchasing goods and ordering services to transferring funds across the country can be done from a small device that fits in our pocket. Ehrlich et al. (2022) highlights that this rapid shift in economic activity creates an opportunity to modernize measurement methods, especially since these transactions generate vast amounts of data that can improve accuracy compared to current processes.

Although the broader RESET project aims to update several indicators, my focus is on the work by Ehrlich et al. (2022) related to the NIPAs, which include GDP. The Bureau of Economic Analysis (BEA) produces these measures—such as productivity and consumer and producer prices—but the system for measuring real and nominal consumer spending is highly decentralized. As an example, the Census Bureau collects retail sales data, while the Bureau of Labor Statistics (BLS) collects price data (see Table 1.1 in Ehrlich et al., 2022). The Census Bureau measures retail sales through monthly and annual surveys and detailed store-level data every five years (the Economic Census). Meanwhile, BLS compiles Consumer Price Index[17] data using information such as expenditure shares from the Consumer Expenditure Survey[18] and a small probabilistic sample of goods through the Commodities and Services Survey.[19] BEA integrates these datasets at a high level of aggregation, facing challenges due to limited product detail between Economic Census years and broad firm classifications in monthly surveys. This structure constrains industry and geographic granularity in key indicators like real GDP, requiring extensive interpolation and extrapolation of other kinds of economic estimates. In short, because NIPAs rely on diverse data sources, inconsistencies often arise when integrating them, as datasets differ in aggregation levels and sampling rates—making micro-level analysis nearly impossible and making industry-level analysis difficult at best.

Addressing current challenges

Modern information technology offers a promising solution to these data collection and integration challenges by replacing the multitude of disparate data sources, agencies, and collection methods into a unified process (Challenges 1, 4, and 5). This ap-

proach enables the simultaneous collection of price and quantity information directly from the original source—the businesses and individuals themselves. Our modern digital infrastructure records each sale of goods and services, often linked to a specific barcode, stock-keeping unit (SKU), or other identifiers that capture individual transactions. This level of detail significantly enhances data quality and accuracy.

If such information could be integrated, Ehrlich et al. (2022) identify three key areas where data collection could improve upon the current system:

1. **Consistency:** price and quantity can be derived from the same observations.
2. **Granularity:** data can be captured at much finer levels across multiple dimensions, such as detailed geographic breakdowns.
3. **Frequency:** time series could, in theory, be constructed at any interval—yearly, monthly, weekly, daily, or even hourly—with minimal lag.

Moreover, these data would represent the full population rather than a sample, dramatically reducing sampling error and minimizing the need for revisions or an extra cleaning step to create a complete dataset (Challenges 2 and 4). This last improvement (frequency) would significantly improve the usability of economic statistics, enabling rapid responses during national economic emergencies or tracking specific goods during seasonal surges such as the holiday shopping period (Challenge 9).

To assess the feasibility, Ehrlich et al. (2022) tested two approaches for measuring prices and real quantities using item-level data: the Unified Price Index approach by Redding and Weinstein (2018, 2020) and an approach applying hedonic methods at scale, similar to Bajari and Benkard (2005). They tested these approaches using two sources of transaction data that are summarized at the item level:

1. **Nielsen retail scanner data:** weekly universal product code (UPC)-level data on expenditures and quantities for millions of products across 35,000 stores, mostly grocery and some mass merchandisers.
2. **NPD Group (NPD, private entity):** private data that covers more than 65,000 general merchandise stores, including online retailers, and focuses on products not in Nielsen's data.

The authors note that the items are defined narrowly at the UPC level in both datasets, so "... dividing sales by units sold gives a good measure of unit price." They state that any change in product attributes generates a new UPC code. This situation means retailers and manufacturers have strong incentives to maintain unique codes for specific products since assigning them is inexpensive.

Persistent and new challenges

Overall, the results from Ehrlich et al. (2022) indicate that using the Unified Price Index methodology (Redding & Weinstein, 2018) combined with hedonics-at-scale methodology for quality adjustment shows strong potential for capturing variation in quality in price indexes using transaction data. However, scalability challenges remain, particularly in implementing hedonics-at-scale. Ehrlich et al. (2022) suggest that machine learning—and likely AI, which advanced significantly after the paper's publication—could support implementation, such as converting text and images into vectors and applying dimensionality reduction methods.

As noted earlier, a version of Challenge 1 arises in standardizing data for capturing the variation in quality in price indexes. Ehrlich et al. (2022) identify three potential approaches for obtaining and standardizing data—each with its own advantages and disadvantages—making it likely that a combination will be required:

1. **Direct feed of transaction-level data:** Federal statistical agencies could receive raw transac-tion-level data directly from firms, but this approach is impractical in the United States due to private companies tending to be reluctant to share granular data and the additional challenges of managing such large volumes of data
2. **Direct feed of (detailed) aggregate measures of price, quantity, and sales via APIs:** A more practical option is for businesses to provide detailed aggre-

gated data (e.g., price, quantity, and sales) through APIs. While this requires substantial upfront investment in infrastructure, it offers long-term benefits of high-quality data at minimal marginal cost.

3. Third-party aggregators: Third-party entities that already collect and aggregate much of the necessary information could supply data to federal statistical agencies, while also enabling firms to respond to requests through these services.

However, more data means more information to process (Challenges 4 and 5). Simply put, greater data availability comes at the cost of curating, storing, and implementing proper security measures to manage access to such rich information. Traditionally, data collection for official statistics places a significant burden on respondents like businesses, requiring them to complete forms based on federal statistical system nomenclature—terminology that often differs from what is commonly used in the private sector. Under the proposed new NIPA framework, this burden shifts from respondents to the federal statistical agencies responsible for producing the official statistics. In other words, federal statistical agencies will face a different processing hurdle (Challenges 1, 4, and 5) of unifying diverse private-sector data at the detailed transactional level to ensure consistency in producing official statistics.

While this shift may increase costs, the trade-off is the potential for substantial improvements in data quality. For example, it reduces reliance on businesses and individuals to accurately complete surveys and forms—tasks that many respondents may not know how to answer correctly or lack the incentive to complete at all—while enabling access to more granular and responsive data (Challenge 2). Also, the situation presents an opportunity to establish universal standards across federal agencies, particularly within the U.S. Census Bureau, BLS, and BEA, which frequently collaborate to produce joint economic data and statistical products.

Beyond the need to standardize processes for collecting and aggregating data, these innovations require a new legal and policy framework for data acquisition, storage, sharing, and transfer. Key considerations include determining fees for obtaining data, ensuring costs remain within budget, and addressing contingencies if companies discontinue services, refuse to sell, or raise prices beyond the federal statistical system's budget. Further, agencies will need to develop the infrastructure that ensures the sharing and transfer of information is safe and secure (Challenges 1, 4, 5, 6, 7, and 8).

Another challenge is workforce readiness (applies to all challenges). Implementing these innovations often requires staff to learn new methods, techniques, programming languages, and technical skills—or hiring new personnel with these capabilities. While this investment adds to short-term costs and runs counter to efforts to reduce expenses, such investments may ultimately prove cost-effective by delivering higher-quality data and more robust economic statistics in the long-term (Ehrlich et al., 2022).

NATIONAL EXPERIMENTAL WELL-BEING STATISTICS

Declining unit and item response rates have increasingly forced federal statistical agencies to rely on weighting adjustments and imputation when estimating income and poverty to maintain population representation and data completeness. At the same time, persistent underreporting of program participation and broader income misreporting raise concerns about the overall quality of these estimates. Administrative records and commercial (third-party) data offer alternative sources of information on income and program participation, though each brings its own challenges related to coverage and bias. Decennial census data, by contrast, provide more complete and precise demographic information than either survey or administrative sources. Integrating administrative and census data with survey data therefore offers a promising strategy for improving the quality of survey-based estimates.

Although individual sources of error in survey estimates have been studied extensively—contributing significantly to our understanding of survey strengths and weaknesses—these studies reveal that different error sources can introduce biases in varying ways. The Na-

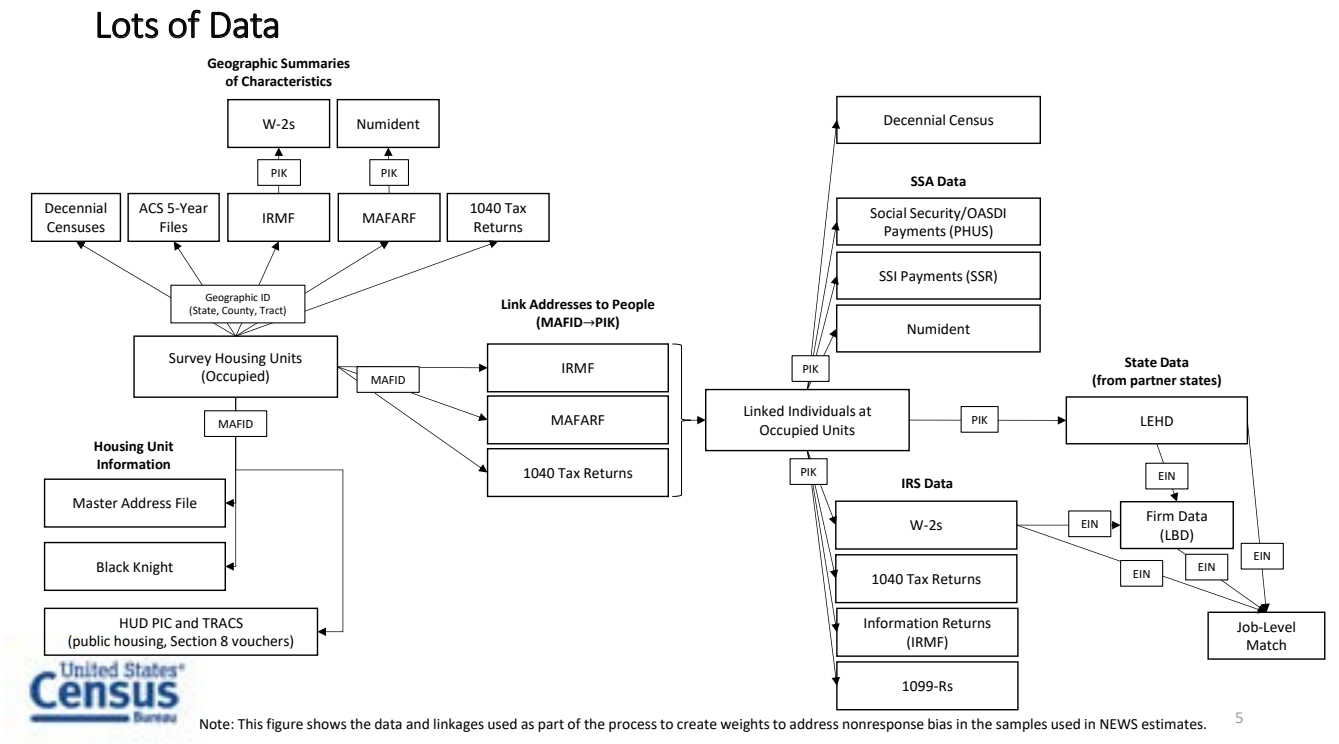
tional Experimental Well-Being Statistics (NEWS) initiative within the U.S. Census Bureau seeks to "rethink how we produce income and resources statistics" by asking: "What is the best possible estimate given all the data currently available at the Census Bureau for a given income or resource statistic?"[20]

Addressing current challenges

This U.S. Census Bureau experimental data product[21] aims to address multiple sources of error simultaneously and, in doing so, create the most accurate estimates of income and poverty currently available (Challenges 1, 2, and 9). In addition to integrating multiple data sources (see Figure 1), NEWS seeks to overcome significant technical and bureaucratic hurdles to produce all versions of the improved income and poverty estimates and make them accessible to a broader audience on their website (Challenges 6, 8, and 11). The Census Bureau emphasizes full transparency in data collection and creation methods (i.e., "transparent, replicable, evidence-based manner"), ensuring that data products include documentation (Challenges 4, 5, and 10).

The Census Bureau NEWS team released its first version in February 2023, limited to 2018 estimates and pre-tax income and poverty measures. They published their work on the NEWS project website and, in September 2023, presented to the Census Scientific Advisory Committee,[22] a group of academic and private-sector experts providing external advice on policy, research, and technical issues across Census Bureau programs. When asked about additional venues for external feedback, the committee recommended several, including the National Tax Association's annual conference and the NBER Conference on Research in Income and Wealth (Challenges 11 and 12).

FIGURE 1



NOTE: Screenshot from the "National Experimental Well-Being Statistics (NEWS); Combining Survey and Administrative Data to Improve Income and Poverty Statistics" presentation, July 2025. Accessed January 7, 2026. https://www2.census.gov/programs-surveys/demo/tables/news/NEWS_detailed_slides.pdf

Since then, the Census Bureau NEWS team has released two new versions (2.0 and 2.5) each incorporating major improvements. Version 2.0 (January 2025) introduced several notable updates, including an expanded definition of income beyond pre-tax money income (Bee et al., 2025). Additional improvements included:

- Estimating federal and state taxes and credits using an in-house microsimulation model (Lin, 2017) which integrates survey responses on income, expenses, and household composition with federal and state tax policies.
- Integrating additional administrative data on means-tested program benefits.
- Updating earnings model to better combine survey and administrative data for estimating the unobserved true earnings distribution.

Version 2.5 (July 2025) further expanded coverage to five additional years, enabled comparisons of income and poverty estimates over time by demographic subgroup, and incorporated administrative data to better model pandemic-related tax credits, such as Economic Impact Payments.[23]

The Census Bureau NEWS team noted that income and poverty measurements were biased by pandemic-related factors, requiring adjustments for nonresponse bias and underreporting. Specifically, improvements included:

- modeling earnings when survey and administrative data contain errors,
- addressing missing and incomplete data (e.g., nonresponse bias in Current Population Survey Annual Social and Economic Supplement (CPS ASEC)[24] and American Community Survey (ACS),[25] survey income nonresponse, incomplete state program data), and
- understanding reporting error (particularly underreporting among 65 and older populations), which varies by income type (especially compared to the BEA numbers; Rothbaum, 2015) and across business cycles.

Persistent and new challenges

A major source of administrative data for NEWS is tax data, including individual tax returns and reports submitted to the IRS by employers, investment firms, and agencies providing income support. However, tax data have limitations in coverage, completeness, and ac-curacy (Challenge 2), such as constructing income for non-filers, underreporting of self-employment income (a key contributor to the "tax gap" (Committee for a Responsible Federal Budget, 2025)), and inaccurate addresses (i.e., reported addresses may not be where filers actually live).

- The Census Bureau NEWS team's efforts to link survey and administrative data still face additional challenges (Challenge 1).
- Data privacy concerns when blending multiple sensitive datasets (Challenge 3) and legal restrictions on data use beyond its intended purpose (Challenge 7). For instance, Title 26 (Internal Revenue Code) restricts how the Census Bureau may use tax data to complement survey estimates and the decennial census.
- Although substantial information about the process, published results, and detailed reports is available, such information does not fully comply with the FAIR principles (Challenge 5).
- Both access and allowed use of program administrative data are also limited (Challenges 6 and 8).
- Uncertain quality of commercial data, which is highly variable and difficult to assess due to proprietary restrictions. This challenge compounds the existing issues associated even with high-quality private data, including budgetary constraints and legal limitations (similar to the issues RESET encounters when using private data) (Challenges 6, 7, and 8).

Future iterations of NEWS aim to improve timeliness by producing estimates more quickly, even if preliminary (Challenges 1 and 4). Additionally, NEWS currently provides regional-level estimates (e.g., Northeast, South) because data sources are aggregated at state, county, and tract levels. The Census Bureau NEWS team plans to leverage ACS and universe-level administrative data to produce more granular state and local estimates (Challenge 2).

COMPREHENSIVE INCOME DATASET PROJECT

With similar motivation as NEWS, the University of Chicago's Comprehensive Income Dataset (CID) Project aims to "address the inaccuracies in our basic understanding of economic well-being in the United States."[26] Simply put, researchers are increasingly turning to administrative data as survey response rates decline and other survey-related challenges increase. Yet, administrative data are far from perfect. They often contain measurement errors and frequently lack key demographic variables that surveys traditionally collect—particularly for vulnerable groups such as adults over age 65. Further, the questions policymakers and communities tend to ask about economic well-being (e.g., "What holes in the safety net remain?") are more varied and nuanced than what most government data systems are designed to answer. As a result, existing economic data and statistics can be biased, incomplete, or poorly aligned with lived experiences.

In response, CID seeks to estimate accurate measures of poverty, deep poverty, and extreme poverty both annually and over multiple decades. CID also strives to assess income disparities across demographic groups and the extent of income inequality in the United States. To accomplish this, the CID team is developing new methods and linking additional, novel data sources that shed light on under covered populations or ones entirely missing from major household surveys—including people experiencing homelessness. Similar to RESET, CID encompasses a broad set of papers addressing diverse economic and public policy questions. I focus on one study by Meyer et al. (2025), which uses CID to estimate that asylum seekers contributed to 60% of the two-year rise in sheltered homelessness in the United States from 2022 to 2024.

Addressing current challenges

The CID's core premise is that no single data source can accurately capture household well-being, but linking multiple sources allows researchers to combine their strengths and compensate for their weaknesses. To achieve this, the project integrates three primary types of data—household surveys, tax records, and administrative program data. Although surveys contain rich demographic detail, tax data provide precise earnings information with broad coverage, and administrative records supply information on benefit receipt often missing from other sources. By connecting these datasets at the individual level through anonymized identifiers, CID produces a more comprehensive and accurate measure of income than any source could provide on its own (Challenges 1 and 2). CID's page also notes that their data will be available publicly to support transparency and reuse (Challenges 10 and 11).

Constructing this unified measure requires substantial methodological innovation. The project uses advanced statistical techniques to impute missing information and guide decisions about how best to merge disparate data sources (Challenges 1 and 2). A key innovation is the use of material hardship indicators—such as housing problems, food insecurity, and mortality patterns—to validate and refine income measurement choices. Currently, the CID has linked four major household surveys with extensive tax data and twelve federal and state administrative datasets, creating what is likely the most complete integrated income dataset in the United States.[27] The project continues to expand data sources and refine measures.

Meyer et al. (2025) found that the Department of Housing and Urban Development (HUD)'s Point in Time (PIT)[28] data indicated a 43% increase in sheltered homelessness from 2022 to 2024—a reversal of a sixteen-year decline. The authors note that media reports, policy discussions, and advocacy campaigns at the time attributed this sharp rise in homelessness to domestic housing conditions, such as worsening affordability and the expiration of pandemic-era eviction protections, often overlooking the role of asylum seekers. To address this gap, Meyer et al. (2025) applied two different approaches: 1. direct estimates from official reports, local tracking systems, and agency correspondence in the four hardest-hit localities; and 2. indirect estimates that assume the Hispanic share of sheltered homelessness would have remained at its 2022 level absent the asylum-seeker influx, attributing the gap between expected and observed counts to asylum seekers.

The HUD data indicated that 79% of this rise was concentrated in New York City, Chicago, Massachusetts, and Denver. For the direct estimates, Meyer et al. (2025) used local administrative data: the NYC Comptroller's Office asylum seeker census, Chicago's official PIT reports (specifically identifying recent migrants), Massachusetts Office of Housing and Livable Communities data on emergency family shelters, and correspondence with the organization that carried out Denver's 2024 PIT counting operation (the Metro Denver Homeless Initiative). They also examined several Department of Homeland Security (DHS) data sources to assess the plausibility of the timing, magnitude, and geographic distribution of the asylum seeker increase on homeless counts (Challenges 1 and 2).

Meyer et al. (2025) used these sources to generate direct estimates of the asylum-seeker share of the 2024 sheltered PIT counts in New York City, Chicago, Massachusetts, and metropolitan Denver. In Massachusetts, they use statewide rather than city-level estimates because the state's centralized shelter placement system distributes asylum seekers across multiple jurisdictions. The state housing office provided a direct estimate of the homeless asylum-seeker population by first analyzing households' immigration status using primary language from surveys conducted after the 2024 PIT and then applying those proportions to the 2024 PIT language distribution to estimate the number of asylum seekers (Challenge 9).

For the indirect method, DHS data indicate that most asylum seekers entering the United States during this period identified as Hispanic. Media reports and Chicago PIT documentation similarly suggest that most migrants entering emergency shelter systems were Hispanic, with the notable exception of Massachusetts, which has a substantial population of homeless Haitian asylum seekers. Meyer et al. (2025) assume that, without the increase of asylum-seekers, the proportion of the sheltered homeless population who are Hispanic would remained at the 2022 level—an assumption supported by the stability of this share between 2016 and 2022 (Challenges 1 and 2).

The direct estimates found that asylum seekers accounted for about 62.2% of the national increase in sheltered homelessness between 2022 and 2024. The indirect estimates, which assume that Hispanic share of the homelessness population would remain constant in absence of the asylum seeker increase, found that asylum seekers accounted for about 59% increase in sheltered homelessness. The authors note that despite the difference in the two estimates, the methods place the contribution of asylum seekers within a reasonably tight range of estimates, with discrepancies that appear to be consistent with known limitations of their methodologies. Overall, Meyer et al. (2025) asserts that asylum seekers are likely the primary driver for the increased number of sheltered homelessness.

Persistent and new challenges

The use and ongoing updates to CID in Meyer et al. (2025) still face several limitations. The authors note that their direct measurement approach does not capture homeless asylum seekers outside the four localities they examined, which likely leads to an undercount of the total number of asylum seekers nationwide. For the indirect approach, they did not incorporate non-Hispanic asylum seekers outside of Chicago and Massachusetts. This omission is important to know given the diversity of the asylum seeker populations in other locations, such as New York City. This gap in their analysis is largely due to limited demographic information on asylum seekers in New York City and other places, which likely caused their indirect estimation method to understate the national impact of asylum seekers on homelessness counts (Challenges 1, 2, and 9).

Both direct and indirect estimation methodologies in Meyer et al. (2025) also rely on the assumption that the influx of asylum seekers did not substantially displace other individuals who otherwise would have entered shelters. If displacement occurred, the estimates would understate what sheltered homelessness would have been absent of the asylum seeker increase. Although available evidence suggests no major shift from sheltered to unsheltered homelessness—unsheltered counts have continued their gradual long-term trajectory—some people may have been diverted into doubled-up living situations or other precarious

housing arrangements instead of accessing shelters (Challenges 1, 2, 9, and 12).

Finally, although asylum seekers account for the majority of the recent rise in sheltered homelessness, the other 40% increase remains unexplained and warrants further investigation, including well-known factors such as "... rising housing costs in major cities, the expiration of pandemic-era eviction protections and supportive services, and complex interactions between new immigrant populations and local housing markets" (Meyer et al., 2025). The authors also note that the 17% rise in unsheltered homelessness between 2022 and 2024 appears to reflect a continuation of a longer-term upward trend that predates the recent asylum seeker influx (Challenges 2, 9, and 12).

SAFE DATA TECHNOLOGIES PROJECT

SOI collects and curates an enormously valuable amount of tax data that supports research on the effects of tax policies and broader empirical questions, such as income inequality. Although only a limited number of government analysts and approved researchers can access the underlying microdata, SOI releases a public use file (PUF) for external researchers and data users. Increasingly, however, the PUF has become difficult to protect with traditional statistical data privacy methods and privacy-enhancing technologies. The growing availability of personal information across public and private datasets, combined with rising computational power, has created unprecedented reidentification and privacy risks.

The Safe Data Technologies Project,[29] a collaboration among the Urban Institute, SOI, and other research institutions, is working to "create a modern data-access system that expands the use of tax data for research purposes while also protecting those data from growing threats of attacks on public data releases" (Burman et al., 2024). The project focuses on improving the PUF through synthetic data generation and on developing another tier of data access between public and secure environments. This new tier would rely on an automated validation server that allows authorized researchers to submit statistical programs to be executed on confidential tax data, with carefully calibrated noise added to the outputs to preserve privacy.

Addressing current challenges

The PUF, a sample of tax returns modified using legacy statistical data privacy techniques, has become increasingly difficult to protect under current computational and information environments (Challenges 1 and 3). In other words, the growing availability of auxiliary personal data and rapid advances in computational methods mean that traditional statistical data privacy approaches can no longer ensure confidentiality without severely degrading statistical validity. As a result, the current process in creating the PUF cannot safely include many policy-relevant variables, such as state of residence or business activity, which substantially limits its usefulness for evidence-based public policymaking (Challenges 2 and 9).

To address these challenges, SOI is developing a fully synthetic PUF that will replace the traditional file. Because synthetic data imitate a confidential dataset and do not contain actual tax return records, the new PUF can include more variables while still preserving privacy and serve as a practical tool for researchers to test and debug statistical code before submitting it to a secure validation environment. Burman et al. (2024) describes the new system as organized into multiple tiers designed to balance data utility with privacy protection (Challenges 1, 2, 3, and 6).

- **Public tabulations**, which is SOI's longstanding program of publicly available tabulations and reports, as well as establishing a Data Services team to produce custom tabulations.
- **Synthetic PUF**, which is designed to replicate key statistical properties of the underlying microdata, supporting most descriptive statistics and some microsimulation applications, but estimates may be less reliable for complex or highly granular statistical queries.
- **Validation server**, where authorized users develop and debug code using the synthetic PUF, then submit programs to a remote access system that executes them on confidential tax data; output include calibrated noise to minimize disclosure risk.
- **Restricted microdata access** under SOI's existing Joint Statistical Research Program. Only a limited number of trusted researchers are granted such access, and all work is conducted under SOI supervision.

Access to all tiers will be granted based on researcher qualifications, research needs, and compliance with statutory requirements under the Internal Revenue Code.

The innovation comes in the development of second and third tiers that expands the accessibility to administrative tax data. For the second tier, the authors explain that the synthesis approach offers several advantages over traditional disclosure-control methods. Synthetic data can represent certain variables more accurately than the traditional PUF, particularly at the tails of the income distribution, and can safely include information previously suppressed for confidentiality such as state of residence, detailed business income, capital losses before limitation, and select information-return data. As SOI expands its synthetic data capabilities, the lag between the availability of confidential data and public release is expected to shrink, increasing the relevance of the data for policy analysis (Challenges 2, 3, and 9).

For the third tier, Burman et al. (2024) describes an automated validation server that provides another secure layer that enables researchers to run analyses on confidential data without directly viewing the underlying, confidential records. Researchers first develop and test their code using synthetic data, which has the same structure as the confidential data, then apply for access to the server. Once approved by an administrator (e.g., SOI staff), they submit their programs to be executed on the confidential dataset. The server returns results—such as regression coefficients, summary statistics, or tabulations—with carefully calibrated noise added to protect privacy (Challenge 3).

The validation server model has the potential to transform how researchers access government administrative data. Traditional pathways—such as downloading PUFs, working as government contractors, or obtaining in-person access at secure facilities—are often slow, burdensome, and geographically restrictive. In contrast, a remotely accessible automated server could streamline the application and approval process while reducing the need for intensive staff review at federal agencies. By eliminating the requirement to travel to IRS offices or secure data centers, the validation server would expand access to researchers regardless of their institutional resources or location, enabling more timely and equitable use of tax data for evidence building (Challenges 6 and 8).

Throughout the project, the Safe Data Technologies team has published numerous papers, blog posts, book chapters, and other written materials; released open-source code on GitHub and developed an R package on CRAN;[30] and presented their work in multiple venues—including conferences in statistics, economics, and public policy, as well as an invite-only scientific advisory board. These activities aim to share knowledge about the tools under development, promote transparency and reproducibility, and solicit feedback from the research community (Challenges 9, 10, 11, and 12).

Persistent and new challenges

Even as synthetic data and validation-server technologies expand access to tax information, they still face several challenges. As with the traditional PUF, analyses conducted on synthetic data can yield statistically invalid results when used beyond their intended purposes. The synthetic PUF aims to support accurate tabulations, simple correlations, and most microsimulation analyses, but it cannot reliably reproduce features such as kinks or discontinuities because the synthesis process inherently smooths these patterns. Evaluating data quality remains a challenge because no single metric captures the utility of synthetic data across all applications. Instead, researchers must rely on multiple use-case–specific assessments, such as comparisons of means and correlations, confidence-interval overlap tests for regression estimates, and consistency with policy-relevant microsimulation outputs (Challenges 9 and 10).

Some data types present more fundamental difficulties. High-resolution geographic identifiers, panel data tracking individuals over time, and corporate income tax records are especially vulnerable to disclosure risks and therefore require much heavier noise infusion, often rendering them less reliable for research purposes. These constraints illustrate the broader privacy-utility trade-off: Stronger safeguards inevitably

reduce accuracy and vice-versa. Synthetic data may still be valuable as "dummy files" or provide the overall structure and range of values of the data without any meaningful results or relationships preserved in the data for code development, even when too noisy for substantive analysis, but they cannot fully substitute confidential data in all contexts (Challenge 3).

Burman et al. (2024) also mentions challenges arise for the validation server. Although the prototype supports a broad set of analyses, researchers remain limited to models compatible with the privacy-preserving algorithm, and certain weighted tabulations or multistage estimators may not yet be feasible. The use of a privacy budget introduces another constraint, where each user must decide how to allocate their finite allowance of information leakage across analyses. At times, this may mean trading precision for quantity of certain statistics. While this mechanism is essential to prevent misuse—such as repeated queries designed to reverse-engineer confidential values—it also requires researchers and data users to plan analyses more deliberately, limits exploratory data analysis, and forces them to understand what the budget actually means in obtaining their statistical outputs. These constraints, while necessary to protect taxpayer privacy, represent ongoing challenges in building a secure, scalable, and analytically useful system for modern administrative data access (Challenges 9, 10, 11, and 12).

Where do we go from here?

The four examples highlighted in this paper offer reasons for hope in the future of economic data and statistics and demonstrate how we can produce better federal economic data that serve the public good. At the same time, these examples point to areas where further improvement is needed and suggest meaningful next steps for the future of federal economic data and statistics.

AI REVOLUTION: THE PROMISE AND LIMITATIONS OF AI

Across the federal statistical system, AI adoption by agencies has so far been mixed, with some researchers excited by its potential to expedite processes while others have concerns about gaps in data governance and protections. These are precisely the reasons why AI cannot be ignored. As Johnson (2026) argues, the federal statistical system cannot overlook AI's potential for "data cleaning, editing, and imputation needed to make administrative data suitable for statistical purposes." This mirrors the insights of Ehrlich et al. (2022) on the RESET project, which emphasizes that machine learning (a subset of AI) can help implement state-of-the-art methods that are otherwise difficult to scale or operationalize, even when these methods are more time-intensive. However, if AI technologies are to be meaningfully integrated, proper storage, metadata standards, and documentation must be prioritized (Challenges 4 and 5).

As noted in Challenge 5, true readiness for the AI era requires ensuring that underlying data and infrastructure are actually AI-ready. George E. P. Box's well-known statement from 1976 that "all models are wrong, but some are useful," remains relevant fifty years later. We cannot blindly trust AI systems to deliver accurate or context-appropriate answers. The quality of underlying data profoundly shapes model performance, and human oversight remains essential to ensure these technologies are used correctly, effectively, and ethically. The NEWS and CID examples illustrate how answering even seemingly simple questions about economic well-being is far more complex than it appears (Challenges 9, 10, 11, and 12).

COST EFFICIENCIES: THE TRADE-OFFS ON EARLY INVESTMENT FOR LONG-TERM REWARDS

One goal of this paper is to explore how the federal statistical system can produce economic data and statistics more efficiently and cheaper through recent innovations. However, the examples discussed

demonstrate that achieving these efficiencies often requires substantial upfront investment. While these initial costs may seem at odds with the aim of reducing spending, periodic updates to IT infrastructure are essential regardless of whether the goal is to improve economic data and statistics. Such updates minimize security risks, increase efficiency (e.g., through hardware that delivers similar computing power at lower energy costs), and allow agencies to retire legacy systems that consume staff time and require costly maintenance across hundreds of outdated programs. Similarly, investing in the professional development of current and new staff—particularly by expanding technical skills—almost certainly improves the efficiency and accuracy of data production and increases staff retention. For example, greater capacity in modern coding practices, state-of-the-art methodologies, and the identification of redundancies or infrastructure gaps contributes directly to higher-quality data and statistics (all challenges ultimately require better infrastructure and investment in staff).

EDUCATION AND COMMUNICATION: ESSENTIAL BUT OFTEN UNDERVALUED

Alongside infrastructure, education and communication across the broader federal statistics and economics communities remain essential. Although nearly every field uses data, very few K–12 or postsecondary programs teach data governance or prepare students to navigate the complexities of privacy, security, and ethical use of data. Even if a person is not actively using data for their profession, our information is constantly being recorded and used without us knowing (Bowen and Williams, 2025a).

Despite the research community and broader population benefiting from learning about data and statistics, resources on these topics are also severely limited. To illustrate the scarcity of guidance and resources, consider two hypothetical scenarios:[31]

Scenario 1	**Scenario 2**
Suppose you had a dataset which contained records of individuals, including demographics such as their age, their sex, and their race. Suppose also that this data contained more in-depth personal information, such as financial records, health status, or political opinions. Finally, suppose that you wanted to glean relevant insights from this data using machine learning or artificial intelligence. What methods would you use? What best practices would you ensure you followed? Where would you seek information to help guide you in this process? For the last question, chances are you can readily think of books, blogs, videos, or other materials suited to different technical skill levels and backgrounds.	Suppose you had a dataset which contained records of individuals, including demographics such as their age, their sex, and their race. Suppose also that this data contained in-depth personal information, such as financial records, health status, or political opinions. Finally, suppose you believed that other people should have access to this dataset, so that they could use this information. Using your knowledge prior to reading this paper, how would you share it? What information would you preserve in the data? What information would you consider too sensitive to share? Where would you seek information on the best practices to go about it?

My guess is that many more readers can readily answer the first scenario than the second without additional research. While data practitioners, researchers, and policymakers frequently use economic data and statistics produced at all levels of government, few receive training in the values of data governance or the challenges that arise across the data life cycle (Challenges 9 and 10).

Even beyond the incentives tied to academic publication, there is limited motivation to create high-quality data products with proper metadata aligned with FAIR principles, provide reproducible code, or invest in documentation. Bowen and Williams (2025b) highlight how current incentives in the research community undervalue the invisible yet essential work performed by data curators. Those in fields that use data should encourage and reward those who create clear, accessible, and actionable educational materials, as well as those who produce data products accompanied by proper metadata—work that benefits everyone who uses data and statistics (Challenges 5 and 10).

SUNSETTING DATA: THE LACK OF INNOVATION OF ARCHIVING AND TERMINATING DATA

Finally, while the four innovative examples highlighted here addressed many of the 15 challenges across the data life cycle, none directly confronted the final three challenges on data archival and termination. When selecting examples for this paper, one colleague created an extensive list of seventeen recent innovations in economic data and statistics, mapping each example to the challenges it addressed. None addressed data archival or termination.

Several factors may help explain this gap. First, several government agencies tend to archive data rather than terminate per laws or other statutory requirements, so innovation in these areas is not required. Second, projects often outlive the people involved in their creation, such as staff changing roles or leaving agencies. For instance, infrastructure managers will often purge hundreds of terabytes of data during system transitions, telling staff to migrate whatever they need. Data and statistics lacking a clear steward or ongoing purpose are easily lost in this transition. Third, many data practitioners are reluctant to delete information, fearing they may need it for reruns or for future, unanticipated projects. Yet storing data indefinitely is expensive. And finally, many institutions may not have clear guidance (or accountability for the guidance) on the archival or termination of data. Generally, the field needs broader awareness and discussion about developing and maintaining policies and procedures that persist beyond individual staff members and adapt as the data ecosystem evolves. Such policies should include systematic checks on whether data should continue to be archived, whether they can be terminated, and, if terminated, which metadata should be retained to document their prior existence. As highlighted in Challenge 15, preserving persistent metadata or summary indicators is essential to maintain accountability and historical context.

As outlined throughout this paper, producing high-quality federal economic statistics requires more than technical expertise. The federal statistical system must have a holistic approach that spans the entire data life cycle and is grounded in the core data governance values of accuracy, accessibility, privacy, and usability. The challenges outlined in this paper underscore the complexity of balancing these values across collection, storage, sharing, analysis, dissemination, and archival. At the same time, the innovations highlighted in the four examples—RESET, NEWS, CID, and Safe Data Technologies—demonstrate that progress is possible when agencies invest in modern infrastructure, adopt adaptive policies, and foster collaboration across disciplines.

However, not all proposed innovations will address the challenges or core values, nor will they be faultless or without new issues to address. Moving forward, sustained commitment to better understanding AI's role in the federal statistical system, modernizing the data infrastructure, developing and investing in workforce skills and educational resources, and upholding responsible data governance will be essential to ensure that federal economic statistics remain robust, trusted, and responsive to the evolving needs of society.

References

Bee, A., Creamer, J., Mitchell, J., Mittag, N., Rothbaum, J., Sanders, C., Schmidt, L., & Unrath, M. (2025). National Experimental Well-Being Statistics: Technical Note on Methodological Changes for Version 2. SEHSD Working Paper FY2025-01, U.S. Census Bureau. sehsd-wp2025-01.pdf.

Benkard, C. L., & Bajari, P. (2005). Hedonic price indexes with unobserved product characteristics, and application to personal computers. Journal of Business & Economic Statistics, 23(1), 61-75.

Bowen, CMK., Bryant, V., Burman, L., Czajka, J., Khitatrakun, S., MacDonald, G., & McClelland, R. Mucciolo, L., Pickens, M., Ueyama, K., Williams, A. R., Wissoker, D., & Zwiefel, N. (2022a). Synthetic individual income tax data: Methodology, utility, and privacy implications. International Conference on Privacy in Statistical Databases, pp. 191-204.

Bowen, CMK., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Mucciolo, L., Pickens, M., & Williams, A.R. (2022b). Synthetic individual income tax data: promises and challenges. National Tax Journal 75(4); 767-790.Bowen, CMK., & Williams, A. R. (2025a). A day in the life with federal government data. Association of Public Data Users. https://apdu.org/?p=5814113

Bowen, CMK., & Williams, A. R. (2025b). Two ideas to support and reward data collectors. Association of Public Data Users. https://apdu.org/?p=5814137

Burman, L., Johnson, B., Bryant, V. L., MacDonald, G., & McClelland, R. (2024). Protecting Privacy and Expanding Access in a Modern Administrative Tax Data System. National Tax Journal, 77(4), 927-947.

Chicago Beyond. (2019). Why am I always being researched? A guidebook for community organizations, researchers, and funders to help us get from insufficient understanding to more authentic truth. Chicago: Chicago Beyond Equity Series.

Committee for a Responsible Federal Budget. (2025). Primer: Understanding the tax gap. https://www.crfb.org/blogs/primer-understanding-tax-gap

Ehrlich, G., Haltiwanger, J. C., Jarmin, R. S., Johnson, D., & Shapiro, M. D. (2022). Reengineering Key National Economic Indicators. Big Data for Twenty-First-Century Economic Statistics, 79, 25.

Johnson, B. (2026). New data sources to improve federal statistics. Association of Public Data Users. https://apdu.org/?p=5814254

Lin, Daniel. 2017. "Methods and Assumptions of the CPS ASEC Tax Model." U.S. Census Bureau SEHSD Working Paper #2022-18Lo Wang, H. (2018). Some Japanese-Americans wrongfully imprisoned during WWII oppose census question. NPR. https://www.npr.org/2018/12/26/636107892/some-japanese-americans-wrongfully-imprisoned-during-wwii-oppose-census-question

Moulton, S., & Luong, J. (2023). Dollars and demographics: How census data shapes federal funding distribution (Project On Government Oversight). https://docs.pogo.org/report/2023/POGO_Federal-Funds-Geographically-Directed-by-Census-Data.pdf

Meyer, B. D., Wyse, A., & Williams, D. (2025). Asylum Seekers and the Rise in Homelessness (No. w33655). National Bureau of Economic Research.

National Academies of Sciences, Engineering, and Medicine. (2015). The Bicentennial Census: New Directions for Methodology in 1990: 30th Anniversary Edition. Washington, DC: The National Academies Press.

Nguyen, D. (2025). ICE made expansive request for taxpayer data amid IRS pushback. Politico. https://www.politico.com/news/2025/10/30/ice-made-expansive-request-for-taxpayer-data-amid-irs-pushback-00630312

Redding, S. J., & Weinstein, D. E. (2018). Measuring Aggregate Price Indexes with Demand Shocks: Theory and Evidence for CES Preferences. NBER Working Paper.

Redding, S. J., & Weinstein, D. E. (2020). Measuring aggregate price indices with taste shocks: Theory and evidence for CES preferences. The Quarterly Journal of Economics, 135(1), 503-560.

Rothbaum, J. L. (2015). Comparing income aggregates: How do the CPS and ACS match the National Income

and Product Accounts, 2007–2012 (SEHSD Working Paper No. 2015-01). U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2015/demo/SEHSD-WP2015-01.pdf

Thessen, A. E., Woodburn, M., Koureas, D., Paul, D., Conlon, M., Shorthouse, D. P., & Ramdeen, S. (2019). Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. Data Science Journal, 18(1), 54.

Endnotes

1 "Census in the Constitution." Accessed January 7, 2026. https://www.census.gov/programs-surveys/decennial-census/about/census-constitution.html

2 "The 2020 Census and Confidentiality." Accessed January 7, 2026. https://www.census.gov/content/dam/Census/library/factsheets/2019/comm/2020-confidentiality-factsheet.pdf

3 "Federal Statistical Research Data Centers." Accessed January 7, 2026. https://www.census.gov/about/adrm/fsrdc.html

4 "ICPSR." Accessed January 7, 2026. https://www.icpsr.umich.edu/sites/icpsr/home

5 "Official 1950 Census Website." Accessed January 7, 2026. https://1950census.archives.gov/

6 "Federal Data Management: Issues and Challenges in the Use of Data Standards." Accessed January 7, 2026. https://www.congress.gov/crs-product/R48053

7 "H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2018." Accessed January 7, 2026. https://www.congress.gov/bill/115th-congress/house-bill/4174

8 "FAIR Principles." Accessed January 7, 2026. https://www.go-fair.org/fair-principles/

9 "Ensuring A National Policy Framework For Artificial Intelligence." Accessed January 7, 2026. https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

10 "Federal Statistical Research Data Centers." Accessed January 7, 2026. https://www.census.gov/about/adrm/fsrdc.html

11 "Effects of the COVID-19 Pandemic on Employment, Earnings, and Professional Engagement: New Insights from the 2021 National Survey of College Graduates." Accessed January 7, 2026. https://ncses.nsf.gov/pubs/nsf23307

12 "Student Loan Debt Statistics in US 2025 | Key Facts." Accessed February 27, 2026. https://theworlddata.com/student-loan-debt-statistics-in-us/

13 "If you have federal student loan debt, here's what experts want you to know." Accessed February 27, 2026. https://www.pbs.org/newshour/nation/if-you-have-federal-student-loan-debt-heres-what-experts-want-you-to-know

14 "Right to erasure ('right to be forgotten')." Accessed January 7, 2026. https://gdpr-info.eu/art-17-gdpr/

15 "California Consumer Privacy Act." Accessed January 7, 2026. https://oag.ca.gov/privacy/ccpa

16 "RESET: Re-Engineering Statistics using Economic Transactions." Accessed January 7, 2026. https://ebp-projects.isr.umich.edu/RESET/

17 "Consumer Price Index." Accessed January 7, 2026. https://www.bls.gov/cpi/

18 "Consumer Expenditure Surveys." Accessed January 7, 2026. https://www.bls.gov/cex/

19 "Consumer Price Index Collection Modes," which have been discontinued as of December 18, 2025. Accessed January 7, 2026. https://www.bls.gov/cpi/tables/collection-modes.htm

20 "National Experimental Well-Being Statistics (NEWS)." Accessed January 7, 2026. https://www.census.gov/data/experimental-data-products/national-experimental-wellbeing-statistics.html

21 "Experimental data products are innovative statistical products created using new data sources or methodologies that benefit data users in the absence of other relevant products. We are seeking feedback from data users and stakeholders on the quality and usefulness of these new products." Accessed January 7, 2026. https://www.census.gov/data/experimental-data-products.html

22 "Census Scientific Advisory Committee (CSAC)." Accessed January 7, 2026. https://www.census.gov/about/cac/sac.html

23 "Economic impact payments." Accessed January 7, 2026. https://www.irs.gov/coronavirus/economic-impact-payments

24 "Current Population Survey Annual Social and Economic Supplement." Accessed January 7, 2026. https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html

25 "American Community Survey." Accessed January 7, 2026. https://www.census.gov/programs-surveys/acs.html

26 "Comprehensive Income Dataset Project Mission." Accessed January 7, 2026. https://cid.harris.uchicago.edu/mission/

27 "Building the CID," has a table of Data Sources Linked or Planned to be Linked to the Comprehensive Income Dataset. Accessed on February 12, 2026. https://cid.harris.uchicago.edu/building-the-cid/.

28 "Point-in-Time Count and Housing Inventory Count." Accessed January 7, 2026. https://www.hudexchange.info/programs/hdx/pit-hic/#2025-pit-count-and-hic-guidance

29 "Safe Data Technologies Project." Accessed January 7, 2026. https://www.urban.org/projects/safe-data-technologies

30 "tidysynthesis: A Common API for Synthesizing Data." Accessed January 7, 2026. https://cran.r-project.org/web/packages/tidysynthesis/index.html

31 These examples are drawn from previous work by the author - Snoke, Joshua, and Claire McKay Bowen. 2020. "How Statisticians Should Grapple with Privacy in a Changing Data Landscape." CHANCE 33 (4): 6–13.

BROOKINGS

1775 Massachusetts Ave NW
Washington, DC 20036
(202) 797-6000
www.brookings.edu