# Synthetic Data for the Nebraska Statewide Workforce & Educational Reporting System

Jeremy Seeman, Aaron R. Williams, and Claire McKay Bowen

*January 2025*

## KEY TAKEAWAYS

Synthetic data, when deployed properly, can provide **stronger disclosure risk protections than those offered by traditional de-identification methods like suppression.**

Most traditional privacy laws focus on **regulating exact, de-identified statistics** instead of the randomized statistics produced by synthetic data.

Unlike traditional privacy law, synthetic data methods treat disclosure risk as a **spectrum to be navigated,** offering more flexibility to data curators.

Synthetic data deployments can **best meet privacy compliance needs when paired with other privacy and security interventions,** like data use agreements (DUAs) and tiered access models.

## OVERVIEW

The Nebraska Statewide Workforce & Educational Reporting System (NSWERS) coordinates data sharing, processing, and dissemination efforts across the Nebraska public school systems, Nebraska community colleges, the University of Nebraska system, the Nebraska Department of Labor, and other statewide partners. Expanding NSWERS data access presents many data governance and privacy challenges. In this summary, we introduce *synthetic data*, a privacy-enhancing technology (PET) that can make NSWERS data more accessible while providing privacy protections.

Synthetic data refers to datasets designed to imitate confidential datasets while limiting information about individuals. Synthetic datasets can be generated by modeling relationships between variables within the confidential data and sampling new "synthetic" records from these models. Doing so provides additional privacy protections beyond traditional de-identification techniques, such as suppression, while minimizing the barriers to share data and therefore reducing administrative burden on NSWERS staff and data users.

Different techniques for performing these modeling and sampling steps help *data curators* (individuals or entities responsible for the safekeeping of the confidential data) navigate the trade-off between *disclosure risk* (our ability to infer information about individual records in the confidential data) and *data utility* (the ability to use synthetic data in the same manner as the confidential data). Once generated, synthetic datasets are evaluated to assess whether a synthetic dataset successfully navigates this trade-off. We introduce different kinds of disclosure risk and utility metrics and their relevance to NSWERS-specific data-sharing scenarios.

As a state longitudinal data system (SLDS), NSWERS has many possible use cases for synthetic data throughout the data lifecycle. We provide example use cases for synthetic data that would improve the efficiency of NSWERS data requests, program evaluation requests, and contractor projects, all while providing stronger privacy protections than more traditional data-access alternatives.

We at the Urban Institute have partnered with the Massive Data Institute at Georgetown University to provide training and technical assistance to NSWERS on generating and evaluating synthetic data. First, our opensource synthetic data–generation and evaluation software has been developed with new features motivated by feedback from the NSWERS team. Second, our "office hours" style technical assistance has provided NSWERS opportunities to become self-sufficient in developing and deploying synthetic data. Overall, these efforts will help NSWERS realize the potential of synthetic data for safely making Nebraska's data more accessible.

## BACKGROUND

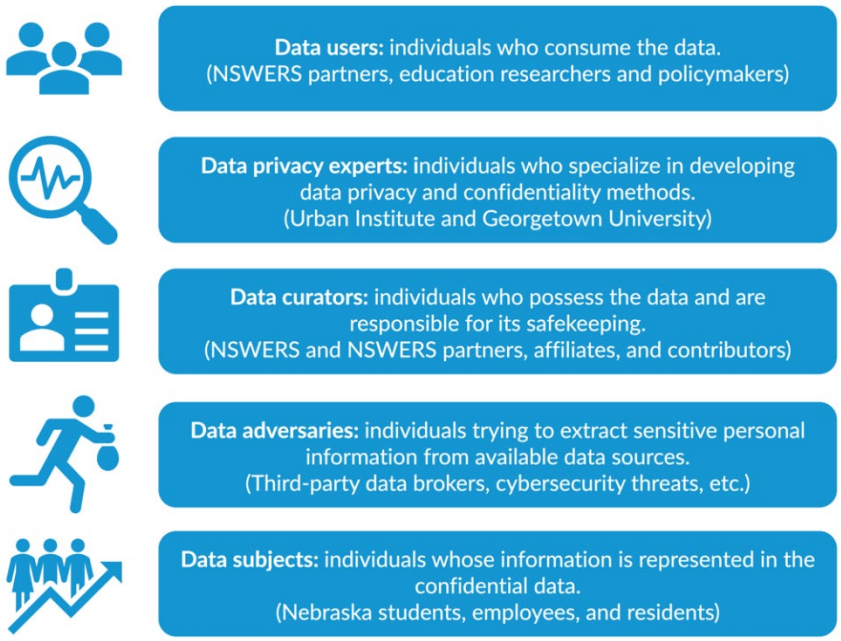
SLDSs aim to make critical education, training, and workforce data accessible to researchers and policymakers. NSWERS is an SLDS that coordinates data sharing, processing, and dissemination efforts across the Nebraska public school systems, Nebraska community colleges, the University of Nebraska system, the Nebraska Department of Labor, and other statewide partners. Consolidating and disseminating these datasets pose significant data governance challenges, especially concerning data privacy. Personal education and workforce data are highly sensitive and frequently subject to specialized data privacy laws like the Family Educational Rights and Privacy Act (FERPA), making the data harder to share with potential data users, such as analysts, researchers, planners, and decisionmakers.

Data privacy concerns are frequently addressed via *de-identification* methods that assess whether certain statistics are safe to release or not (Hundepool et al. 2012). These methods tend to involve procedural rules and processes, such as removing personally identifying information (PII) fields like names and addresses or suppressing (i.e., preventing the release of) statistics that describe a small number of individuals in the dataset. Many privacy laws and their legal interpretations have language that explicitly references these traditional techniques (e.g., suppressing small counts for FERPA compliance). Although these techniques can be helpful in releasing the data publicly, they can be restrictive in two critical ways:

1. **Traditional de-identification techniques can still leave confidential data vulnerable to privacy risks.** Even when de-identification techniques like suppression have been applied, it is possible to infer information about individuals contained in the confidential dataset. In other words, removing PII fields is no longer sufficient to guarantee privacy protection (Ohm 2009). A recent study showed that data available on EdX (an online course platform) that was protected with traditional de-identification techniques still allowed for the exact reconstruction of individual student records, posing concerns for FERPA compliance (Cohen 2022).
2. **Traditional de-identification techniques are often not sufficient to expand access to critical datasets.** For many data products, like those produced at NSWERS, more sensitive or granular data that might aid researchers and policymakers is often inaccessible or otherwise difficult to access due to privacy concerns. For example, many data sources available within NSWERS are not available to the public, instead requiring that prospective data users complete a research proposal application. Because of these limitations, de-identification techniques can either overly restrict access to useful data or provide insufficient disclosure risk protections (Dwork and Naor 2010). These limitations suggest finding alternative approaches to data sharing that offer more flexibility in addressing privacy concerns, which is precisely where synthetic data comes into play.

FIGURE 1

**Participants in the NSWERS Data Ecosystem**




URBAN INSTITUTE

**Source:** Authors' illustration.

## UNDERSTANDING SYNTHETIC DATA

To understand synthetic data, we should first understand the problem synthetic data are trying to solve. Figure 1 lists the relevant participants in the NSWERS data ecosystem. The population of Nebraska's students, employees, and residents make up our *data subjects*, those whose information is represented within NSWERS systems. NSWERS staff and its partners, affiliates, and data contributors are *data curators*, individuals who possess and manage safe access to data describing our data subjects. Much of NSWERS datasets are confidential, meaning they contain sensitive information that cannot be publicly disclosed; therefore, NSWERS must decide how it provides access to *data users*, individuals who consume the data, such as education researchers. A challenge is that some data users could be *data adversaries*, those aiming to extract sensitive personal information from existing data sources, such as third-party data brokers.

Our goal as data privacy experts is to create an alternative process to access information based on confidential data without direct access to the confidential data itself. Note that any dataset containing information individuals prefer to keep private can be classified as sensitive or confidential, even if that information contains no PII fields. Synthetic data provides an alternate method for accessing data that approximates the statistical properties of an existing confidential dataset.

### *What Are Synthetic Datasets?*

A synthetic dataset is built to imitate a confidential dataset while limiting information about individual records. Synthetic data can refer to either the methodology used to produce synthetic datasets or the synthetic data output itself (Raghunathan 2021). In this context, imitating a confidential dataset means the synthetic data have similar statistical properties to the confidential data. Data processing and analysis should produce similar outputs when

**KEY DATA PRIVACY TERMINOLOGY**

**Confidential data:** a dataset that contains personal or sensitive information that is not publicly accessible

**Synthetic data:** a dataset designed to imitate a confidential dataset while limiting information about individual records in the confidential dataset.

**Privacy-enhancing technology (PET):** a computational or algorithmic process or method used to limit the unintended leakage of personal information in data processing tasks

**Disclosure risk:** the risks of unintentionally disclosing personal information about contributors to a confidential dataset by inferring information about them from published statistics or data

**Data utility:** the ability for synthetic data to mimic the properties of a confidential dataset, either in general or in specific data processing tasks

using either the confidential or synthetic dataset. Synthetic data have higher utility or usability when the outputs from the confidential and synthetic datasets are more similar.

Limiting information about individual records makes it harder for a data adversary to infer information about the data subjects in the confidential dataset. A data adversary should have difficulties inferring whether a data subject contributed their information to the underlying confidential dataset used to generate the synthetic data.

***Purposes of Synthetic Data***

All data sharing poses some privacy risks, regardless of how the data are released. Simply put, releasing information based on confidential data cannot perfectly preserve both privacy and utility. A perfectly secure dataset is one that is not accessible, which is useless. Conversely, a perfectly useful dataset is one that is released without any privacy protection.

A synthetic dataset is an example of a PET, a computational or algorithmic process used to limit the unintended leakage of personal information in data-processing tasks. Different PETs address different kinds of privacy harms, which may emerge from system design, algorithmic processes, access controls, or other technical aspects of data-processing systems. For example, secure multiparty computation is a PET that enables computation on multiple combined data sources while preventing third-party intermediaries from accessing individual data sources. In general, PETs refer to computational frameworks for providing privacy protections. The degree or type of protection guarantees varies significantly based on how the practical implementation choices balance disclosure risk with usefulness.

Synthetic data specifically aims to mitigate *disclosure risks*, or the risks of unintentionally disclosing personal or sensitive information about data subjects that can be inferred from published statistics or data based on a confidential dataset. Using synthetic data affords the following:

- **Greater disclosure risk protections.** Even when suppressing or removing data fields containing PII, providing access to confidential data without further modification can still enable users to infer personal information about individuals in the confidential dataset (Dwork et al. 2017). Synthetic data involves randomization based on a statistical model to provide greater protection, making such inferences much harder, thus limiting disclosure risks.
- **Expanded data accessibility.** Existing processes for getting access to sensitive datasets can be burdensome. Alternatively, synthetic data can be more easily shared than the confidential data, often requiring only a modest data-use agreement instead of more extensive processes. This reduces the administrative burden for both data curators and data users.

## HOW SYNTHETIC DATASETS ARE GENERATED

### *Overview of the Synthetic Data–Generation Process*

Effective synthetic data deployment involves multiple steps, outlined in figure 2. First, in step 0, data curators work with data users and subjects to define use cases for synthetic data. This critical step ensures that synthetic data meets the needs of prospective data users, which may vary significantly. Some data users may only need a few key variables, such as sociodemographic outcomes and individual NSWERS outcomes like graduation rates. Other users may need more granular information, such as detailed course enrollment information for individual students. As part of this process, data curators can identify *thresholds*, or goals for the success of a particular synthetic data task. For example, program evaluators may want to produce statistics that differ by a relative error of no more than 10 percent from those based on the confidential data.

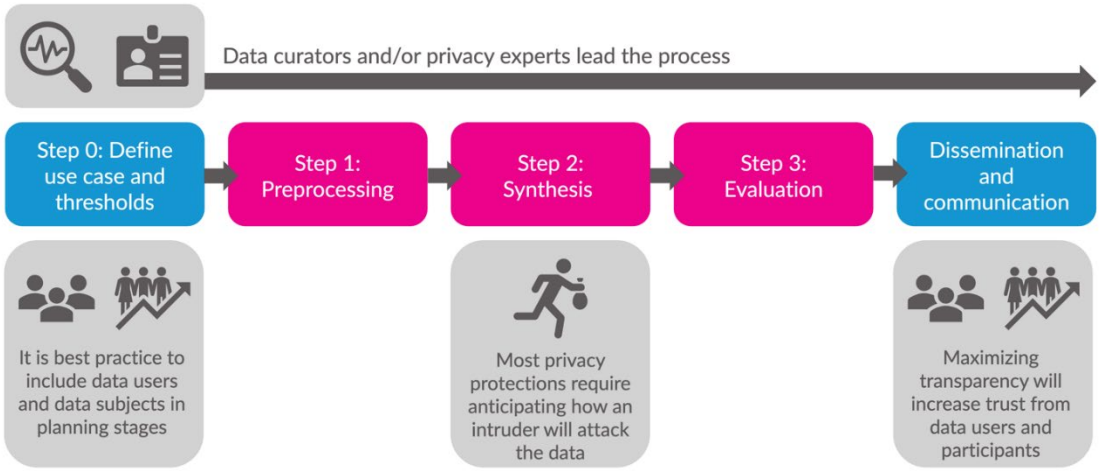
Next, in step 1, data curators or privacy experts work together to preprocess the data. The goal of this stage is to design and create a *gold-standard dataset (GSDS)*, or a dataset that contains all the possible information data users need to perform their analyses. Synthetic data always aims to imitate one specific GSDS, but there are many possible ways to design and produce GSDSs. For example, some education researchers may be interested in student-level synthetic data (one record per student), whereas others may be interested in course-level synthetic data (one record per instance of a particular course).

In the next two steps, we generate and evaluate our synthetic data. In step 2, we use statistical or algorithmic modeling techniques to generate synthetic data based on our GSDS (this is described in more detail in the next subsection). In step 3, we evaluate our synthetic data to ensure it meets our privacy goals and our use case goals. Steps 2 and 3 are often performed iteratively; one might evaluate multiple synthetic data configurations or try new synthetic data configurations motivated by evaluation results.

Finally, in the last step, synthetic data are disseminated to potential data users. Here it is critical to make sure the data users are well-informed about how the synthetic data were generated, how the synthetic data were evaluated, and how the synthetic data should be used in practice. For example, if a synthetic dataset is sufficiently accurate for producing two-way contingency tables but not for complex nonparametric modeling, data users need to know this information to responsibly use their synthetic data.

FIGURE 2

**Synthetic Data Process Overview**




URBAN INSTITUTE

**Source:** Authors' illustration.

### *Synthetic Data–Generation Methods*

In general, synthetic data are generated in two stages: first, the data curators or data privacy experts build a statistical model to capture relationships between variables that appear in the confidential dataset. Next, those same experts randomly sample synthetic records or values from these models.

Often, these two stages occur more than once. Many synthetic data methods, including the ones used by Urban Institute's software, uses what's known as *fully conditional synthesis*. This type of synthesis means new variables (i.e., tabular data columns) are iteratively added one by one to construct synthetic data. For example, suppose we have already synthesized two demographic variables, like gender and age, and we want to synthesize a new variable, whether someone was ever a transfer student. We would first model the proportions of transfer students based on the two demographic variables (e.g., the proportion of males ages 21 to 24 that are transfer students). Then, we would randomly sample new values for each synthetic record's transfer student status according to this model.

Crucially, synthetic data only allow users to perform statistical analyses that are supported by the synthesis process. If a synthetic data user is interested in investigating a particular aspect of the confidential data distribution, that aspect must be accounted for as part of the modeling process. Continuing the previous example, suppose we modeled the proportion of transfer students based only on age, not age and gender. This would prevent a synthetic data user for investigating the three-way relationship between gender, age, and transfer-student status because the synthetic data model assumes one's gender is not associated with their transfer-student status.

### *Modeling Considerations*

All synthetic data's effectiveness hinges on successful model building. The following considerations have the biggest impact on both disclosure risks and utility; see (Drechsler et al. 2024) for more methodological details.

- **Synthesis inputs and ordering.** When using fully conditional synthesis, not all variables can be effectively modeled using the same predictor variables. For example, if a model for graduation rate relies on predictors that fail to capture statistically meaningful differences in graduation rates, the resulting synthetic data may be lower quality or have less utility. This situation is why it is important to synthesize variables in an order that captures increasingly complex relationships. Doing so improves the quality of variables synthesized later in the process.
- **Model complexity.** There are countless different methods for modeling the relationships between variables. Generic model fitting methods (sometimes known as nonparametric, empirical or "black box" methods) allow for the construction of more complex, flexible models than simpler methods (such as parametric or generative models like regressions). For example, the relationship between degree programs and income is quite complex and can be more effectively modeled using generic methods. Although more complex models can produce higher utility synthetic data, they can also produce models that overfit to the confidential data, leaking information about data subjects in the process.

## HOW SYNTHETIC DATA ARE EVALUATED

Synthetic data must always be evaluated to ensure both sufficient utility and disclosure risk protection prior to being publicly released or being accessed by data users. Evaluating synthetic data helps data curators like NSWERS communicate about how synthetic should (or should not) be used and the privacy risks involved with making synthetic data more widely available.

### *Data Utility*

Data utility evaluations measure how well synthetic mimics the structural and statistical properties of the confidential data. Evaluations typically involve computing metrics that compare statistics based on the confidential data versus the synthetic data.

There are two main classes of utility metrics (Snoke et al. 2018). First, *general utility* (or global utility) metrics measure how distinguishable two entire datasets are from one another. For example, suppose we have a synthetic dataset based on confidential data that contains categorical demographic information about students, such as their county, gender, and race. We could count the number of unique students in each category for each dataset and compute the average difference in counts between the synthetic and confidential datasets.

Second, *specific utility* metrics measure how similar synthetic and confidential data perform on the same predefined task or analysis. For example, suppose we wanted to estimate a regression coefficient or parameter that describes the association between transfer student status and graduation rates. We could apply the regression analysis to compute the parameter estimate on the synthetic and confidential datasets separately. We then compare the signs, magnitudes, uncertainties, and so on, of the synthetic and confidential estimates.

### *Disclosure Risk*

Disclosure risk refers to the ability to infer information about data subjects in confidential datasets using publicly released statistics, synthetic datasets, or other resources that are based on the confidential data. Disclosure risk metrics depend heavily on assumptions about what information might be available about how synthetic data were generated or the underlying confidential dataset.

Disclosure risk metrics capture different kinds of inference about data subjects:

- **Membership inference** measures how well someone can infer the presence or absence of data subjects within the confidential dataset. For example, can a prospective data user infer whether someone they know contributed their information to NSWERS in the synthetic dataset?
- **Attribute inference** measures how well someone can infer information about data subjects within the confidential dataset. For example, can a prospective data user infer a data subject's ACT score using the synthetic data? Note that given sufficient background information, successful attribute inferences can also reveal specific data utility. In other words, some synthetic data models that produce higher utility will capture relationships between variables that exist within the broader population, not just the specific confidential sample. For example, if graduation rates differ by age in the general Nebraska population, one could use age to estimate graduation rates for individuals within NSWERS data.

Disclosure risk metrics also come in two flavors:

- **Empirical metrics** describe how effectively and practically someone can use synthetic data to infer information about data subjects in the confidential dataset. For example, a data adversary could generate multiple synthetic data samples and measure the frequency of more unique individuals within the synthetic data and their relationship to the confidential data.
- **Formal metrics** describe how effectively the process that generates synthetic data provides protection. For example, PETs like differential privacy can measure the amount of randomized noise is injected into the synthetic data-generation process (Dwork et al. 2014).

## PRIVACY POLICY AND SYNTHETIC DATA

### *Privacy Compliance and Synthetic Data*

Synthetic data techniques have been studied in academic and government statistics for over 30 years. However, the regulatory status of synthetic data for satisfying privacy law remains open, primarily because federal privacy laws in the US normally apply to exact statistical data releases and not randomized data releases like those produced by synthetic data methods. Many flagship privacy laws often predate modern computing and data infrastructures that enable both greater privacy risks and better protective measures. For example, FERPA was established in 1974,

predating almost all modern developments in data-sharing infrastructure, machine learning, and artificial intelligence. Additionally, social science data (including but not limited to education and workforce data) is more widely available in auxiliary or external data releases, like social media data. These compounding factors have made grappling with definitions of personal information particularly challenging (Wu 2013).

Major privacy laws typically rely on absolute rules about the definition of PII and their disclosure. For example, FERPA lists certain fields, such as student names, addresses, ID numbers as PII; for other combinations of fields, FERPA relies on a "reasonable person" standard, wherein a "reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, [could] identify the student with reasonable certainty."

Synthetic data methods, as opposed to privacy laws based on absolute rules, recognize that *any* combination of fields could be potentially identifying in the right context. The goal of synthetic data and many other PETs is to recognize disclosure risk as a spectrum and to provide more options to efficiently navigate that spectrum. It remains up to legal expert determination whether synthetic data meets reasonable person standards because there is natural contestation around what constitutes a "reasonable person" and "reasonable certainty." Still, because traditional privacy laws have a narrow focus that synthetic data methodology expands upon, properly implemented synthetic data has the potential to enable both greater data access and disclosure risk protections recognized by legal scholars as "in excess" of existing US privacy laws (Zeide 2015).

Note that the privacy law landscape is rapidly changing at the state level. Many states, following efforts in California and Virginia, are adopting privacy laws that mirror the main provisions in the European Union's General Data Protection Regulation (GDPR). Although these regulations have broader definitions of personal or sensitive information, many questions remain about the role PETs play in satisfying these provisions (Finck and Pallas 2020).

***Using Synthetic Data with Other Data Governance Tools***

Generating synthetic datasets is one of many methods in the data sharing and privacy toolkit, and no single tool can solve all data governance problems. For example, disclosure risk protections like those afforded by synthetic data releases cannot protect against unauthorized confidential data access or misuse. Similarly, some data-processing tasks should not use synthetic data as input substitutes, no matter the quality of the synthetic data, such as probabilistic record linkage (Gkoulalas-Divanis et al. 2021). For these reasons, we recommend pairing synthetic data with other data governance tools, such as the following two approaches.

First, synthetic data functions best as part of a "tiered access" approach to data sharing. In this setting, prospective data users have multiple options for how they access their data, and the approval process required to access data varies at each level or option. Providing tiered options makes it easier for prospective data users to identify the right quantity and quality of data needed for their tasks. For example, prospective data users interested in summary statistics about the NSWERS population could have reduced data access burden, such as submitting a full research proposal application, by gaining access to synthetic data; alternatively, potential data users interested in running more complicated unit-level models or those interested in small subpopulations may need to request complete access to the confidential data. Tiered access approaches help to provide additional safety layers and liability protections for data curators, giving them the confidence to safely expand access to more data.

Second, more prospective data users can leverage synthetic data when they have mechanisms or processes for verifying the similarity of results run on synthetic and confidential data. Such mechanisms can take many forms. Methodological transparency about how synthetic data were generated can help users investigate where differences between the synthetic and confidential data might occur. Sharing data utility metrics with prospective data users can help them understand what tasks are most similar or dissimilar with synthetic data. To provide the most comprehensive support, data curators can provide interactive query systems, like table builders and validation servers, that enable synthetic data users to interactively compare results for more generic analyses.

### *Examples of Synthetic Data*

Synthetic data has been successfully deployed in many government data-sharing contexts. For example, the US Census Bureau has many synthetic data products, including the Synthetic Survey of Income and Program Participation (Benedetto et al. 2013) and the Synthetic Longitudinal Business Database (Kinney et al. 2011). Similarly, the Internal Revenue Service has programs dedicated to developing synthetic data and validation server approaches to expand access to tax data (Bowen et al. 2022). In educational contexts, synthetic data has many possible applications for SLDSs besides NSWERS. For example, the state of Maryland has piloted their own synthetic data solutions for making their education and workforce data more accessible (Goldstein et al. 2020). For examples of synthetic data evaluation, NIST's Collaborative Research Cycle provides resources and example evaluations for synthetic data algorithms used on the American Community Survey (Sen et al. 2024).

## NSWERS USE CASES FOR SYNTHETIC DATA

NSWERS has numerous opportunities to use synthetic data for expanded data access. Currently, NSWERS's research data stores can be accessed through their public "insights" website, containing open web-based reports, and their "insights+" layer, their secure data access portal. Synthetic data offers a third alternative, providing more granular information than that contained in public reports with easier accessibility than requesting insights+ portal access. Below, we outline a few hypothetical scenarios that could help NSWERS researchers.

### *Exploratory Educational Research*

Suppose a group of education researchers at the University of Nebraska Omaha are interested in studying demographic population changes for University of Nebraska Omaha graduates and how that affects whether they pursue work in the state of Nebraska or elsewhere. These researchers want to apply for an Institute of Education Sciences grant to pursue this research, but they only need preliminary or exploratory results to justify the scientific merit of their work for the grant. Moreover, the researchers do not yet know what kinds of analyses might be possible with access to NSWERS's secure data portal, which may prohibit them from providing a sufficiently complete research proposal application to NSWERS.

In this scenario, these researchers would immediately benefit from NSWERS synthetic data access in three ways. First, the researchers could use the synthetic data to determine which NSWERS data analyses could provide promising results. Second, the researchers could provide preliminary results based on synthetic data in their grant applications without going through the full research proposal application process. Finally, should the Institute of Education Sciences grant be awarded, the researchers could then provide a more comprehensive and thorough research proposal application to NSWERS that would reduce the administrative burden for both the researchers and NSWERS. Furthermore, the analysis on the synthetic data could support analyses run directly on the confidential data followed by manual or automated disclosure reviews.

### *Data Contributor Engagement*

Suppose a PK-12 electronic transcript provider is interested in contributing course metadata to NSWERS that enhances existing data describing PK-12 curriculums. The transcript provider aims to develop data-processing tools that can augment the existing NSWERS PK-12 curriculum data before it enters the NSWERS data store. To determine how to design, implement, and validate their system, they would like example data to begin testing their system prior to working with NSWERS.

The transcript provider would benefit from synthetic data by enabling end-to-end integration testing of their systems without immediate access to confidential data. By building and validating their systems on synthetic data, they could develop most of their proposed integrated data-processing system without requiring access to

NSWERS's secure data portal. Furthermore, the transcript provider would not be subject to any development constraints placed on the secure portal.

***Expedited Program Evaluation***

Suppose administrators at Northeast Community College (NCC) would like to assess the effectiveness of a new, experiential technical training program on wages. They want to provide results divided by different sociodemographic subpopulations, but they do not know enough about the availability of wage data for NCC graduates to determine which subpopulations could be adequately evaluated.

NCC could leverage NSWERS synthetic data to estimate the number of program participants with wage data in the NSWERS system. This would enable them to more efficiently and responsibly select which sociodemographic subpopulations to include in their program evaluation request to NSWERS.

## URBAN INSTITUTE AND NSWERS

The Urban Institute and the Massive Data Institute at Georgetown University partnered to provide opensource software and technical assistance to NSWERS as they develop and deploy their own synthetic data solutions. Urban provided two pieces of software: *tidysynthesis*, an R package for building and sampling from synthetic data models and *syntheval*, an R package for evaluating synthetic data. The Urban Institute also provided biweekly office hours and technical assistant during the training period, where we helped NSWERS navigate specific implementation questions while discussing new features we developed for the software, motivated by their feedback. Through our collaboration, the NSWERS staff can self-sufficiently generate synthetic data without granting the Urban Institute or the Massive Data Institute at Georgetown University access to their confidential datasets.

## FREQUENTLY ASKED QUESTIONS

***Is Synthetic Data "Fake Data"?***

The word "synthetic" can evoke negative connotations about data appearing inauthentic or deceptive. Such connotations often deter data curators from adopting synthetic data due to anxieties about institutional or reputational harms from releasing "fake" data products. Although concerns about data quality and responsible data usage are important, these concerns apply to any data product, synthetic or not. Modifications to raw data, such as input error correction, imputations of missing values, or record linkages, result in statistical decisions by a user that affect the analysis. Alternatively, we should view synthesis as another part of the data-making process, where greater transparency about data-production processes can enable more responsible use.

***How Do I Validate or Provide a Validation Process for Synthetic Data?***

Validation processes allow data users to ensure that results produced based on synthetic data are like those that could be hypothetically produced by the confidential data. Different approaches to validation rely on answering a few design questions. First, how would data users interact with the validation process? Some may submit manual requests to data curators, whereas others may rely on automated interfaces, such as validation query systems. Second, what kinds of validation metrics would be surfaced to users? Some validation processes provide exact statistics that quantify precise differences between confidential and synthetic results, whereas others use PETs to protect the validation results themselves. Finally, how might data users use the results of the validation process? Some processes allow for all validation results to be released publicly, but others may be omitted or suppressed due to disclosure risk concerns.

### *How Do I Help Users Effectively and Responsibly Use Synthetic Data?*

New users of synthetic data may be tempted to directly substitute confidential data for synthetic data, but making this substitution without additional considerations could be dangerous in the wrong setting. User education is the most effective tool to enable responsible use and prevent potential misuse. Most importantly, responsible user education starts with explicit disclaimers about how synthetic data was generated and what the synthetic data should and should not be used for in downstream applications. More comprehensive user education can take many forms, including reporting on synthetic data evaluations, trainings, and other learning resources.

### *How Do I Decide What Confidential Datasets to Synthesize?*

Synthetic datasets should always be motivated by use cases where the GSDS (i.e., the data to synthesize) is reasonably inaccessible and the data curators could feasibly synthesize a dataset meeting a critical mass of use-case needs. Many decisions about how to select your GSDS should be driven by use cases; for example, if only a small subset of questionnaire items is regularly used in a survey instrument, it would be wise to prioritize synthesizing only these heavily used questionnaire items.

## GLOSSARY

- **Attribute inference:** the process of inferring information about data subjects within the confidential dataset
- **Confidential data:** a dataset that contains personal or sensitive information that is not publicly accessible
- **Data adversaries:** individuals trying to extract sensitive personal information from available data sources
- **Data curators:** individuals who possess the data and are responsible for its safekeeping
- **Data subjects:** the individuals whose information is represented in the confidential data (e.g., residents of Nebraska)
- **Data users:** individuals who consume the data
- **Data utility:** the ability for synthetic data to mimic the properties of a confidential dataset, either in general or in specific data processing tasks
- **Disclosure risk:** the risks of unintentionally disclosing personal information about contributors to a confidential dataset by inferring information about them from published statistics or data
- **Membership inference:** the process of inferring the presence or absence of data subjects within the confidential dataset
- **Privacy-enhancing technology (PET):** a computational or algorithmic process or method used to limit the unintended leakage of personal information in data processing tasks
- **Synthetic data:** a dataset designed to imitate a confidential dataset while limiting information about individual records in the confidential dataset

## REFERENCES


Benedetto, Gary, Martha Stinson, and John M Abowd. 2013. "The Creation and Use of the SIPP Synthetic Beta." Washington, DC: US Census Bureau.

Bowen, Claire McKay, Victoria L. Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Livia Mucciolo, Madeline Pickens, and Aaron R Williams. 2022. "Synthetic Individual Income Tax Data: Promises and Challenges." *National Tax Journal* 75 (4): 767–90. https://doi.org/10.1086/722094.

Cohen, Aloni. 2022. "Attacks on Deidentification's Defenses." In *31st USENIX Security Symposium (USENIX Security 22)*, 1469–86.

Drechsler, J., D. Kifer, J. Reiter, and A. Slavković. 2024. *Handbook of Sharing Confidential Data: Differential Privacy, Secure Multiparty Computation, and Synthetic Data*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press. https://books.google.com/books?id=PdEdEQAAQBAJ.

Dwork, Cynthia, and Moni Naor. 2010. "On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy." *Journal of Privacy and Confidentiality* 2 (1): 93–107. https://doi.org/10.29012/jpc.v2i1.585.

Dwork, Cynthia, Aaron Roth, and others. 2014. "The Algorithmic Foundations of Differential Privacy." *Found. Trends Theor. Comput. Sci.* 9 (3–4): 211–407. http://dx.doi.org/10.1561/0400000042.

Dwork, Cynthia, Adam Smith, Thomas Steinke, and Jonathan Ullman. 2017. "Exposed! A Survey of Attacks on Private Data." *Annual Review of Statistics and Its Application* 4:61–84. https://doi.org/10.1146/annurev-statistics-060116-054123.

Finck, Michèle, and Frank Pallas. 2020. "They Who Must Not Be Identified—Distinguishing Personal from Non-Personal Data under the GDPR." *International Data Privacy Law* 10 (1): 11–36. https://doi.org/10.1093/idpl/ipz026.

Gkoulalas-Divanis, Aris, Dinusha Vatsalan, Dimitrios Karapiperis, and Murat Kantarcioglu. 2021. "Modern Privacy-Preserving Record Linkage Techniques: An Overview." *IEEE Transactions on Information Forensics and Security* 16:4966–87. https://doi.org/10.1109/TIFS.2021.3114026.

Goldstein, Ross, Michael E Woolley, Laura M Stapleton, Daniel Bonnéry, Mark Lachowicz, Terry V Shaw, Angela K Henneberger, Tessa L Johnson, and Yi Feng. 2020. "Expanding Mlds Data Access and Research Capacity with Synthetic Data Sets." Baltimore, MD: Maryland Longitudinal Data System Center.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. 2012. *Statistical Disclosure Control*. Vol. 2. Wiley New York.

Kinney, Satkartar K, Jerome P Reiter, Arnold P Reznek, Javier Miranda, Ron S Jarmin, and John M Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." *International Statistical Review* 79 (3): 362–84. https://doi.org/10.1111/j.1751-5823.2011.00153.x.

Ohm, Paul. 2009. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA l. Rev.* 57:1701. https://ssrn.com/abstract=1450006.

Raghunathan, Trivellore E. 2021. "Synthetic Data." *Annual Review of Statistics and Its Application* 8 (1): 129–40. https://doi.org/10.1146/annurev-statistics-040720-015536.

Sen, Aniruddha, Christine Task, Dhruv Kapur, Gary Howarth, and Karan Bhagat. 2024. "Diverse Community Data for Benchmarking Data Privacy Algorithms." Advances in Neural Information Processing Systems 36. https://doi.org/10.48550/arXiv.2306.13216.

Snoke, Joshua, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." *Journal of the Royal Statistical Society Series A: Statistics in Society* 181 (3): 663–88. https://doi.org/10.1111/rssa.12358.

Wu, Felix T. 2013. "Defining Privacy and Utility in Data Sets." *U. Colo. L. Rev.* 84:1117. https://dx.doi.org/10.2139/ssrn.2031808.

Zeide, Elana. 2015. "Student Privacy Principles for the Age of Big Data: Moving beyond FERPA and FIPPS." *Drexel L. Rev.* 8:339. https://ssrn.com/abstract=2821837.

## ABOUT THE AUTHORS

**Jeremy Seeman** is a research associate in the Data Governance and Privacy Practice Area at the Urban Institute. His research focuses on technical and policy solutions for responsible data sharing, with an emphasis on deploying privacy-enhancing technologies (PETs) to expand access to sensitive data. Jeremy has worked on PETs methodology and deployments for the US Census Bureau, the National Science Foundation, and the Internal Revenue Service. Outside of the Urban Institute, Jeremy is on the steering committees for the National Institute of Statistical Sciences New Researchers Network and the American Statistical Association Privacy and Confidentiality Interest Group. He is an adjunct professor at the Institute for Social Research and a faculty affiliate of the Center for Ethics, Society, and Computing at the University of Michigan.

**Aaron R. Williams** is a lead data scientist for statistical computing in the Data Governance and Privacy Practice Area at the Urban Institute. He works on data privacy, data imputation, microsimulation modeling, and survey analysis with a focus on income, wealth, tax, and retirement policies. Williams has developed systems for safely releasing administrative data for research, including the Urban-Brookings Tax Policy Center's synthesis of individual tax records and the US Department of Labor's Safe Transfer, Restricted-Use Data Lake. He is a member of the Bureau of Labor Statistics Technical Advisory Committee and an adjunct professor in the McCourt School of Public Policy at Georgetown University.

**Claire McKay Bowen** is a senior fellow and leads the Data Governance and Privacy Practice Area at the Urban Institute. Her research focuses on developing technical and policy solutions to safely expand access to confidential data for advancing evidence-based policymaking. She also has interest in improving science communication and ensuring people are properly represented in data. In 2024, she became an American Statistical Association Fellow "for her significant contributions in the field of statistical data privacy, leadership activities in support of the profession, and commitment to mentoring the next generation of statisticians and data scientists." Further, she is a member of the Census Scientific Advisory Committee and several other data governance and data privacy committees as well as an adjunct professor at Stonehill College.

**ACKNOWLEDGMENTS**

This summary was funded by the Bill and Melinda Gates Foundation via Georgetown University. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples. 